Introduction to cluster analysis and classification: Evaluating clustering

Christophe Biernacki

Introduction to cluster analysis and classification: **Evaluating clustering**

C. Biernacki

Summer School on Clustering, Data Analysis and Visualization of Complex Data
May 21-25 2018, University of Catania, Italy

Evaluating clustering

"Technical" evaluation

$$\hat{z} = f(\boldsymbol{x}, \delta[, \Delta, \text{kernel}, \ldots], K, \text{algo})$$

"User" evaluation

A good clustering result is an end-user useful clustering result

Need always to combine both evaluation points of view

Outline

1 Data factor

2 Dissimilarity factor (and co)

3 Algorithm factor

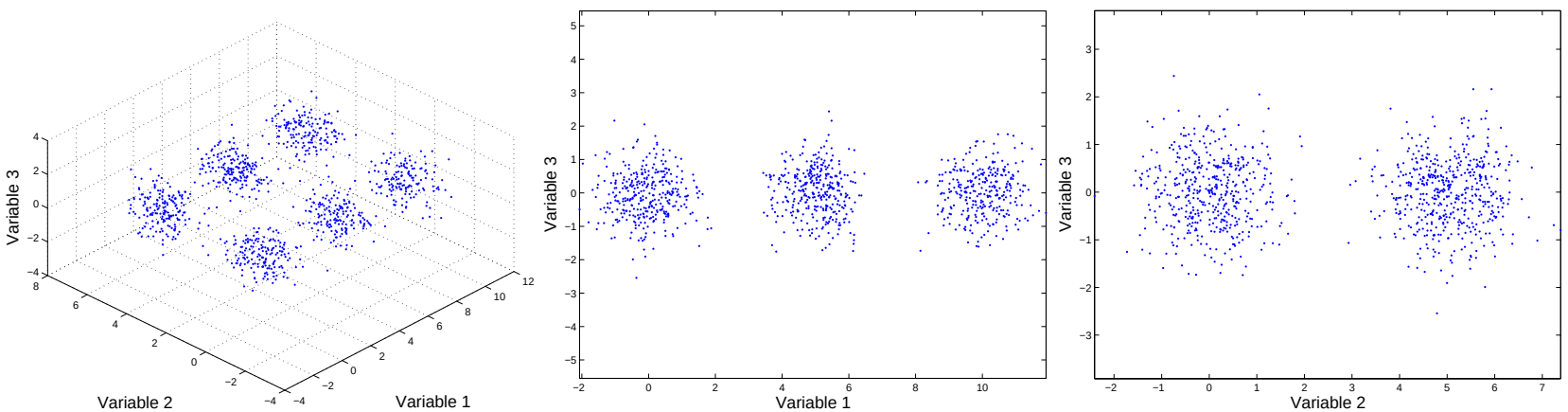
4 Number of clusters factor

5 User factor

6 To go further

The variable effect

- Medicine[1]: diseases may be classified by etiology (cause), pathogenesis (mechanism by which the disease is caused), or by symptom(s). Alternatively, diseases may be classified according to the organ system involved, though this is often complicated since many diseases affect more than one organ.
- And so on...



[1] Nosologie méthodique, dans laquelle les maladies sont rangées par classes, suivant le système de Sydenham, & l'ordre des botanistes, par François Boissier de Sauvages de Lacroix. Paris, Hérissant le fils, 1771

Need to compare partitions: empirical error rate

- Two partitions $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$
- τ: all permutations on $\{1, \ldots, K\}$
- Empirical error rate

$$\text{err}(\boldsymbol{z}, \hat{\boldsymbol{z}}) = \frac{1}{n} \min_{\tau} \sum_{i=1}^{n} \mathbb{I}_{\{z_i = \tau(\hat{z}_i)\}} \in \left\{0, \frac{K-1}{K}\right\}$$

- Partitions are closer when err is small
- Restricted to compare partition with the same number of clusters
- Example

$\boldsymbol{z}$	$\hat{\boldsymbol{z}}$	$\text{err}(\boldsymbol{z}, \hat{\boldsymbol{z}})$
$G_1 = \{a, b, c\}$	$\hat{G}_1 = \{e, f\}$	$\frac{1}{6} \min\{5, 1\} = \frac{1}{6}$
$G_2 = \{d, e, f\}$	$\hat{G}_2 = \{a, b, c, d\}$	

Need to compare partitions: rand index

- Two partitions z and $\hat{z}$
- A measure on basis of agreement vs. disagreement between object pairs
- Not limited to the same number of clusters between partitions
- Rand index [Rand 1971]
  - A: #pairs of elements in x that are in the same subset in z and in the same subset in $\hat{z}$
  - B: #pairs of elements in x that are in different subsets in z and in different subsets in $\hat{z}$
  - C: #pairs of elements in x that are in the same subset in z and in different subsets in $\hat{z}$
  - D: #pairs of elements in x that are in different subsets in z and in the same subset in $\hat{z}$

$$\text{rand}(z, \hat{z}) = \frac{A + B}{A + B + C + D} = \frac{\text{nb. agree}}{\text{nb. agree} + \text{nb. disagree}} \in \{0, 1\}$$

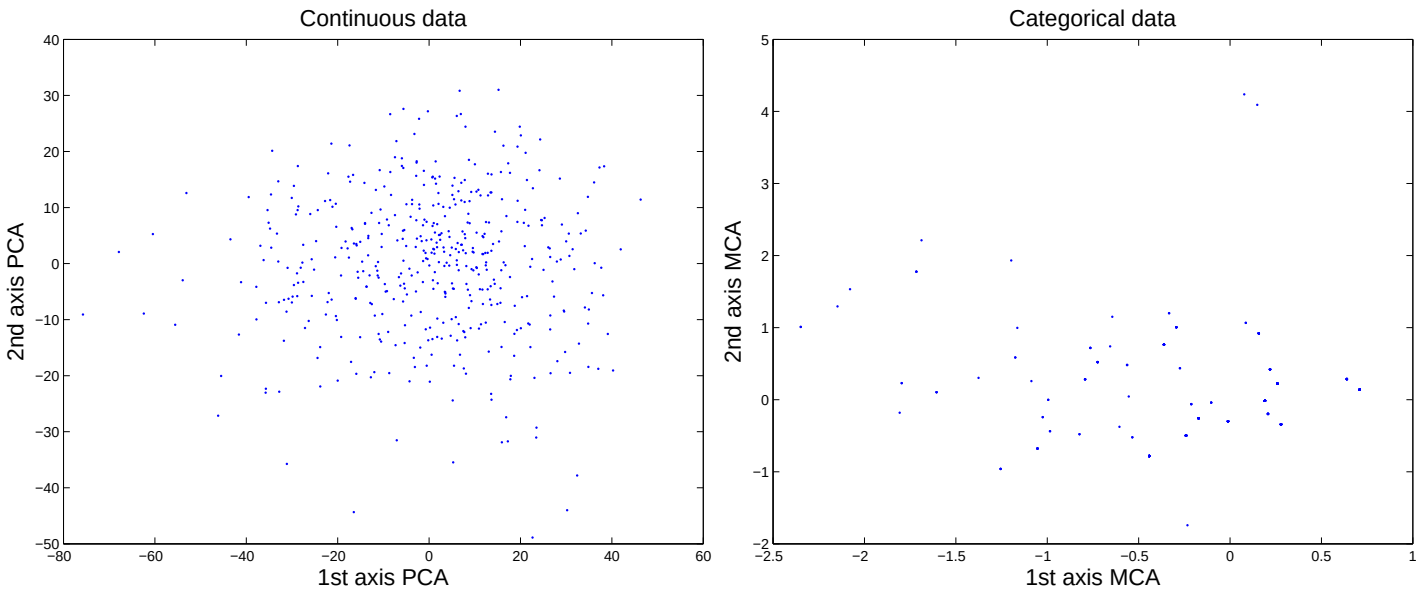
- Partitions are closer when rand is high
- Example

z	$\hat{z}$	intermediate	$\text{rand}(z, \hat{z})$
$G_1 = \{a, b, c\}$	$\hat{G}_1 = \{a, b\}$	$A = 2,\ B = 7$	0.6
$G_2 = \{d, e, f\}$	$\hat{G}_2 = \{c, d, e\}$	$C = 4,\ D = 2$	
	$\hat{G}_3 = \{f\}$		

- Caution: use the adjusted rand index [Hubert and Arabie 1985] to compare $\text{rand}(z, \hat{z})$ and $\text{rand}(z, \tilde{z})$ when $\hat{K} \neq \tilde{K}$

Prostate cancer data: description[2]

- 475 patients from 506 (missing values have been discarded)
- 8 quantitative variables, 4 categorical (some are ordinal) variables
- Two "evident" clusters for medical users: Stage 3 and Stage 4 of cancer



[2] Byar and Green (1980)

Prostate cancer data: variable detail

Covariate	*Abbreviation*	*Number of Levels (if categorical)*
Age	Age	
Weight	Wt	
Performance rating	PF	4
Cardiovascular disease history	HX	2
Systolic Blood pressure	SBP	
Diastolic blood pressure	DBP	
Electrocardiogram code	EKG	7
Serum haemoglobin	HG	
Size of primary tumour	SZ	
Index of tumour stage and histolic grade	SG	
Serum prostatic acid phosphatase	AP	
Bone metastases	BM	2

Prostate cancer data: partition according to retained variables

	quantitative err=9.46%		categorical (raw) err=47.16%		mixing quali/quanti err=8.63%	
	1	2	1	2	1	2
Stage 3	247	26	142	131	252	21
Stage 4	19	183	120	82	20	182

- Partition varies with retained variables as expected
- A general principle: categorical variables less informative than quantitative ones
- However, categorical variables here improve quantitative ones

Prostate cancer data: partition according to recoded variables

	categorical (raw) err=47.16%		categorical (MCA) err=38.95%	
	1	2	1	2
Stage 3	142	131	175	98
Stage 4	120	82	87	115

- MCA is equivalent to recoding categorical variables
- Raw data and MCA data are in a one-to-one mapping (no info. loss)
- It can however drastically impact clustering result
- It open the question of data units/coding to use
- Currently: let the user to choose the unit (prior or posterior choice)
- Next lesson: need formalizing to go further

Prostate cancer data: partition according to missing data

- Use the reduced data set without individuals having missing data ($n = 475$)
- Use the completed data set where missing data are imputed[3] ($n = 506$)
- In both cases, use all mixed variables (not all details at this step, see next lesson)

Data set	completed data	reduced data
err	12.8	8.1

- It is current to have a data "pretreatment" like missing data imputation
- Be careful: it can impact the clustering
- Imputation gives only an estimate data set $\hat{\boldsymbol{x}}$ which is a "deteriorated" data set
- As a consequence it can lead to a "deteriorated" clustering result
- See next lesson to formalize this problem

[3] We use the mice package:http://cran.r-project.org/web/packages/mice/mice.pdf

Stability of a clustering result

- Do not forget that $\hat{\boldsymbol{z}}$ is just an estimate of (a hypothetical true) $\boldsymbol{z}$
- Statistical properties of this estimate should be addressed, as it stability (variance)
- A simple (but computational demanding) attempt:
  - Use bootstrap samples $\boldsymbol{x}^{(b)}$ $(b = 1, \ldots, B)$
  - Obtain bootstrap partitions $\boldsymbol{z}^{(b)}$
  - Deduce for instance confidence regions on centers $\boldsymbol{\mu}$ through related centers $\boldsymbol{\mu}^{(b)}$
  - Be careful to the permutation of labelling!
- See the next lesson for more on the statistical properties (need formalizing)...

Outline

1 Data factor

2 Dissimilarity factor (and co)

3 Algorithm factor

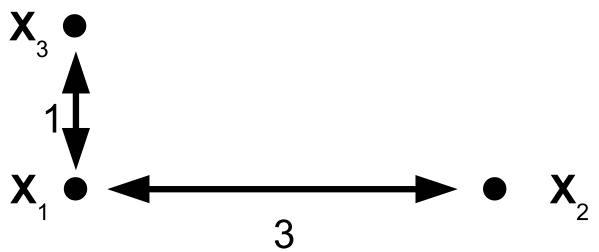
4 Number of clusters factor

5 User factor

6 To go further

Effect of the metric **M** (1/5)

$$\mathcal{X} = \mathbb{R}^2, \mathbf{M} = \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{x}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{x}_2 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \boldsymbol{x}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$



$$\delta_{\mathbf{M}}(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 = (\boldsymbol{x}_1 - \boldsymbol{x}_2)' \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} (\boldsymbol{x}_1 - \boldsymbol{x}_2) = a(x_{21} - x_{11})^2 = 9a$$

$$\delta_{\mathbf{M}}(\boldsymbol{x}_1, \boldsymbol{x}_3)^2 = (\boldsymbol{x}_1 - \boldsymbol{x}_3)' \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} (\boldsymbol{x}_1 - \boldsymbol{x}_3) = (x_{32} - x_{12})^2 = 1$$

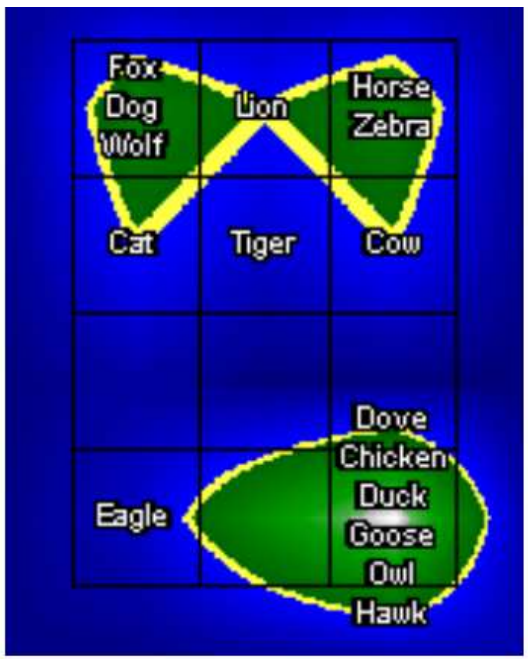
Effect of the metric **M** (2/5)

$$\delta_{\mathbf{M}}(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 \leq \delta_{\mathbf{M}}(\boldsymbol{x}_1, \boldsymbol{x}_3)^2 \Leftrightarrow a \leq \frac{1}{9}$$

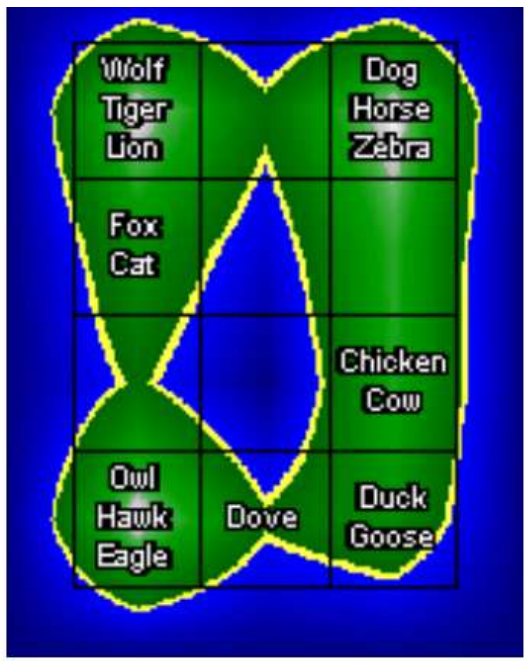
- The distance is impacted by the metric, thus the clustering could be also
- Somewhere the metric is also related to variable selection (try $a = 0 \ldots$)

Effect of the metric **M** (3/5)

- Animals represented by 13 Boolean features related to appearance and activity
- Large weight on the appearance features compared to the activity features: the animals were clustered into mammals vs. birds
- Large weight on the activity features: partitioning predators vs. non-predators
- Both partitions are equally valid, and uncover meaningful structures in the data
- The user has to carefully choose his representation to obtain a desired clustering



(a)

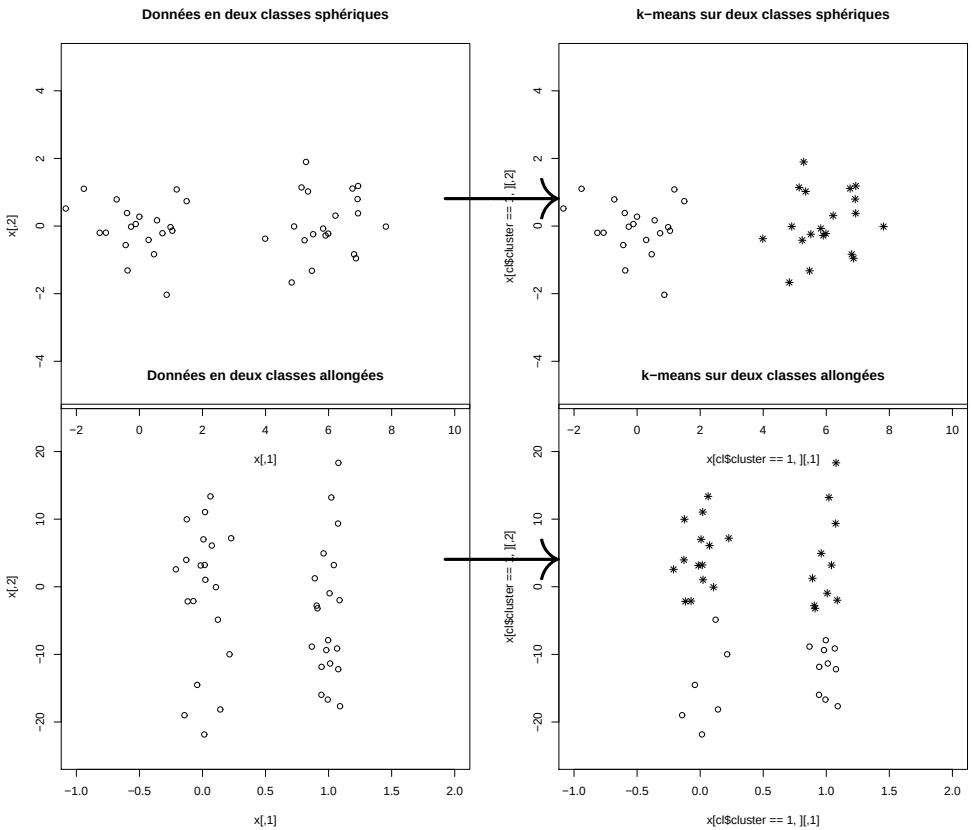


(b)

Figure 6 Different weights on features result in different partitioning of the data. Sixteen animals are represented based on 13 Boolean features related to appearance and activity. (a) partitioning with large weights assigned to appearance based features; (b) a partitioning with large weights assigned to the activity features (figure reproduced from [**Pampalk *et al.* , 2003**]).

Effect of the metric **M** (4/5)

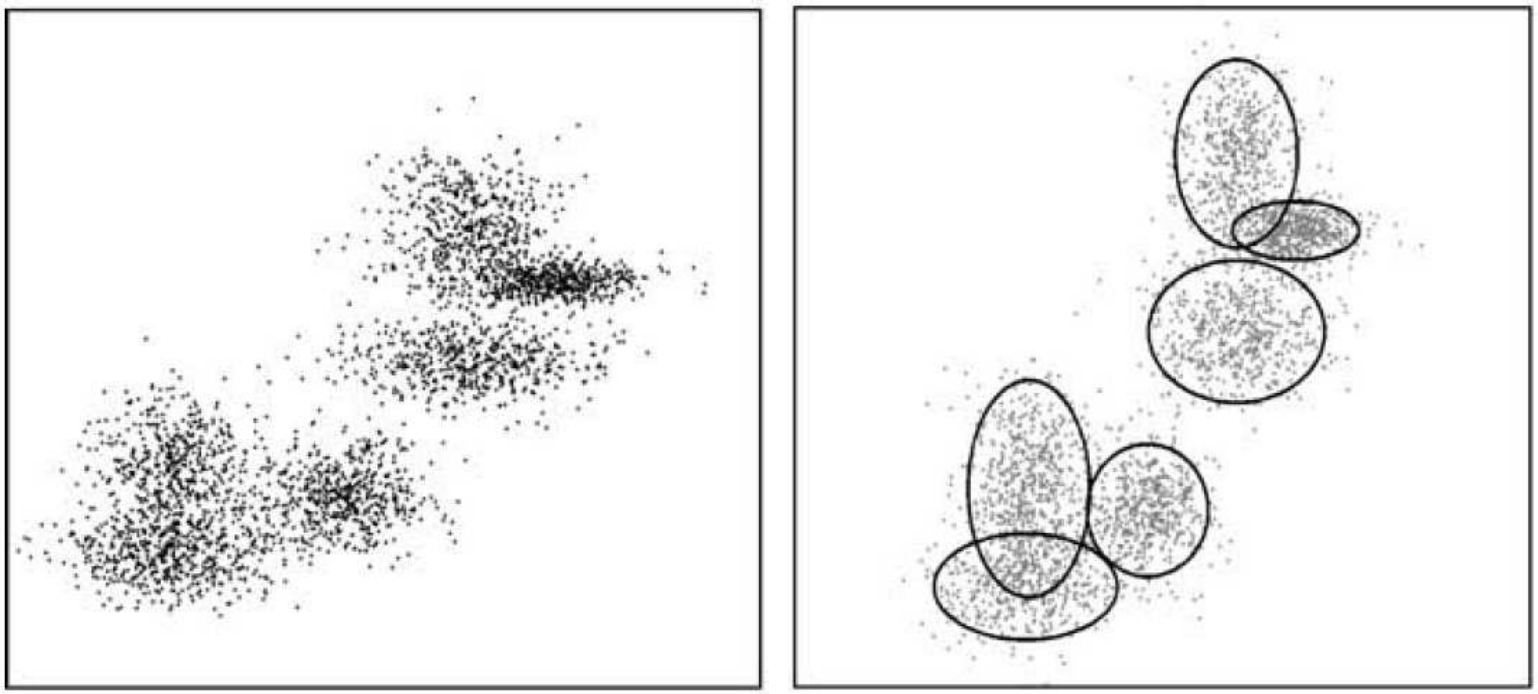
If $\mathbf{M} = \mathbf{I}$...



Alternative: estimate $\mathbf{M}_{(k)}$ by minimizing $W_{\mathbf{M}_{(k)}}(\mathbf{z})$ over $(\mathbf{z}, \mathbf{M}_{(k)})$

Effect of the metric[4] $\mathbf{M}$ (5/5)

Alternative: estimate $\mathbf{M}_{(k)}$ by minimizing $W_{\mathbf{M}_{(k)}}(\mathbf{z})$ over $(\mathbf{z}, \mathbf{M}_{(k)})$



[4] Figures from A.K. Jain (2008). Data Clustering: 50 Years Beyond K-Means.

Effect of the linkage criterion (1/3)

[A. Jain *at al.*. Data Clustering: A Review.]

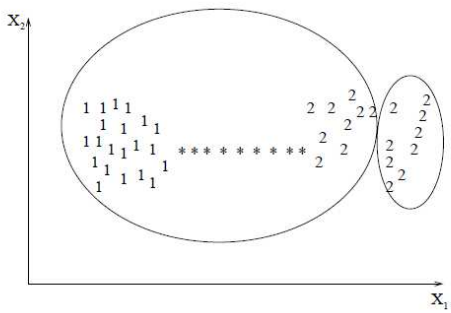


Fig. 12. A single-link clustering of a pattern set containing two classes (1 and 2) connected by a chain of noisy patterns (*).

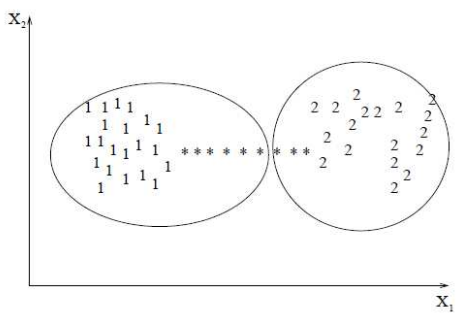


Fig. 13. A complete-link clustering of a pattern set containing two classes (1 and 2) connected by a chain of noisy patterns (*).

Effect of the linkage criterion (2/3)

[P.-N. Tan *at al.* (2005). Introduction to data mining, second edition, Addison-Wesley, Chap.8]

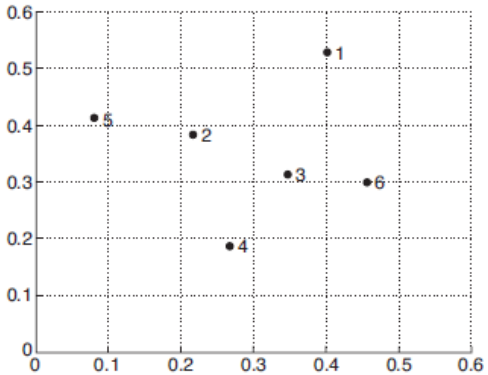


Figure 8.15. Set of 6 two-dimensional points.

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

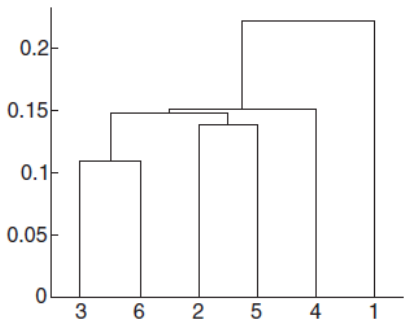
Table 8.3. xy coordinates of 6 points.

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

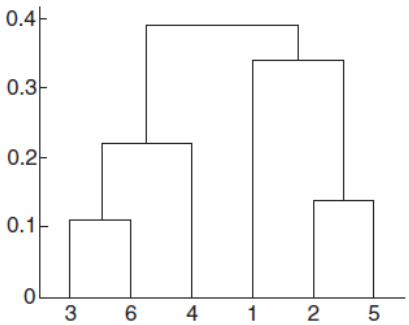
Table 8.4. Euclidean distance matrix for 6 points.

Effect of the linkage criterion (3/3)

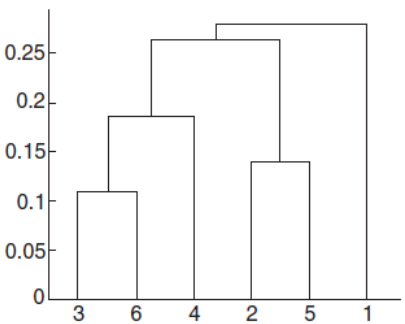
[P.-N. Tan *at al.* (2005). Introduction to data mining, second edition, Addison-Wesley, Chap.8]



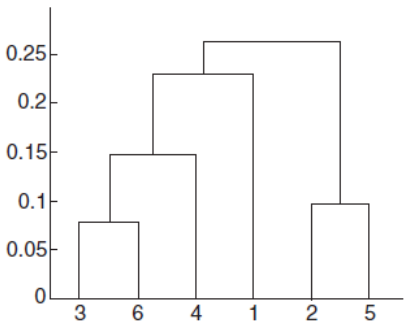
Single link dendrogram.



Complete link dendrogram.



Group average dendrogram.



Ward's dendrogram.

What is "and co"?

Notice also obviously that:

- Kernel clustering result depends on the kernel choice
- Spectral clustering result depends on the Laplacian choice
- ...

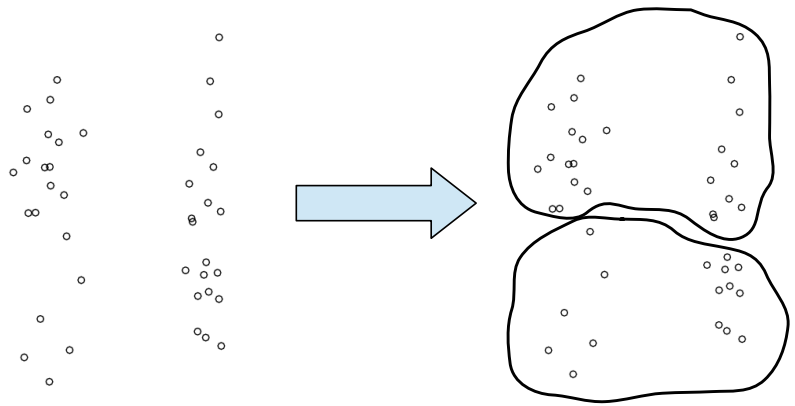
A meaningful way to be less metric (and co) dependent: idea

- Clustering interesting if separated clusters
- If separated clusters, partition less metric dependent
- Thus, the problem is partially reported on choosing K (see later in this lesson)
- It will be an interesting element to be used in next lesson also

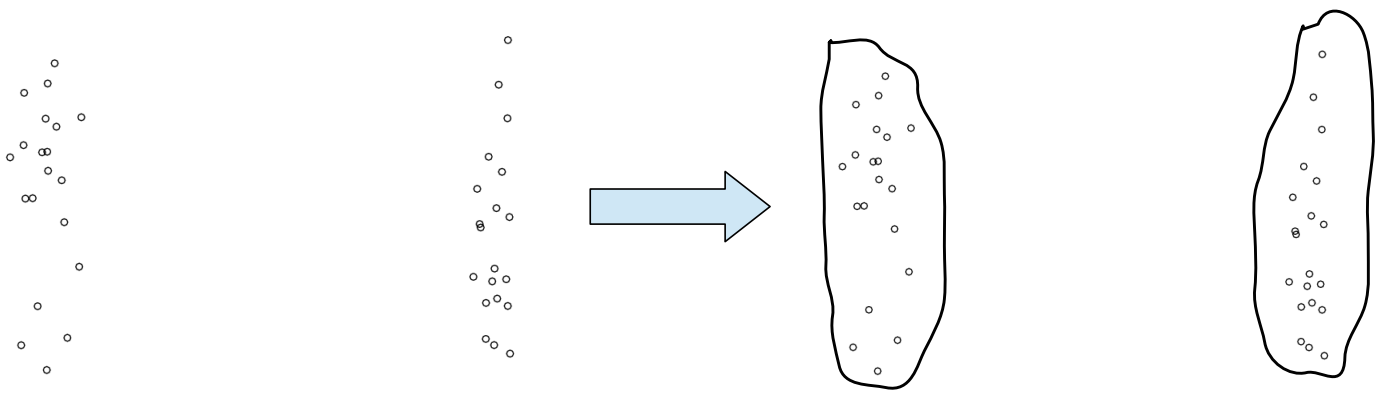
This idea is also applicable for hierarchy, kernel, spectral clustering...

A meaningful way to be less metric (and co) dependent: illustration

- K-means with $\mathbf{M} = \mathbf{I}$
- Not well-separated clusters

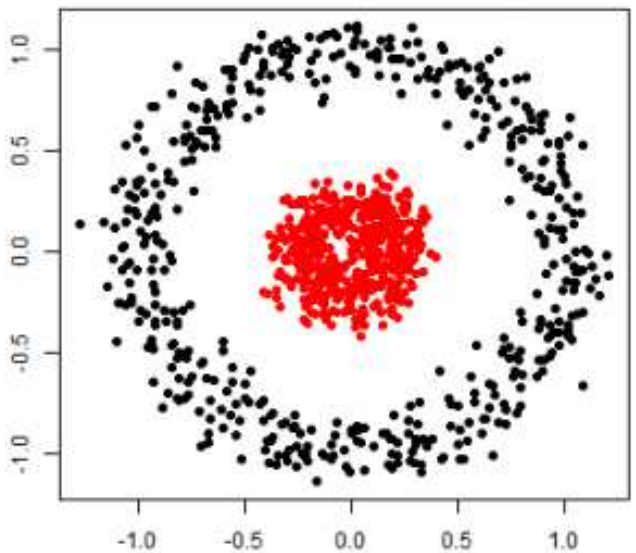


- Well-separated clusters

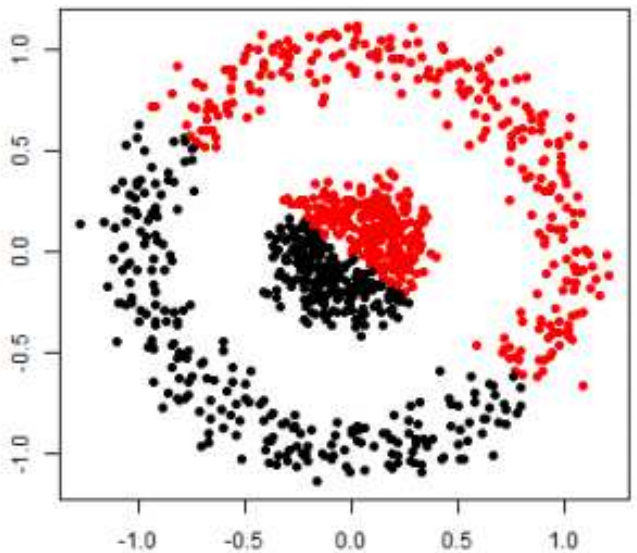


A meaningful way to be less metric (and co) dependent: limit

However, it is not always sufficient. . .



Spectral clustering



K-means clustering

Outline

1 Data factor

2 Dissimilarity factor (and co)

3 Algorithm factor

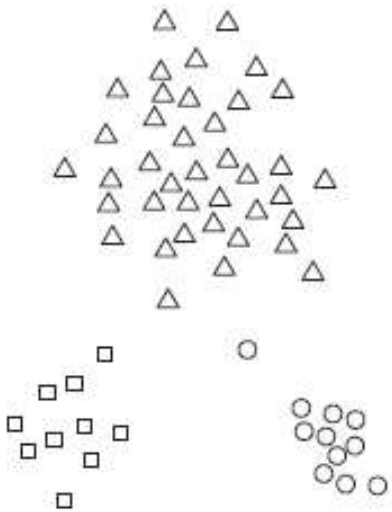
4 Number of clusters factor

5 User factor

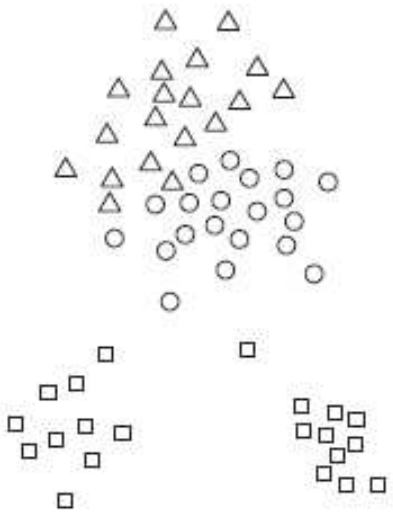
6 To go further

Local maxima with K-means: example

[P.-N. Tan *at al.* (2005). Introduction to data mining, second edition, Addison-Wesley, Chap.8]



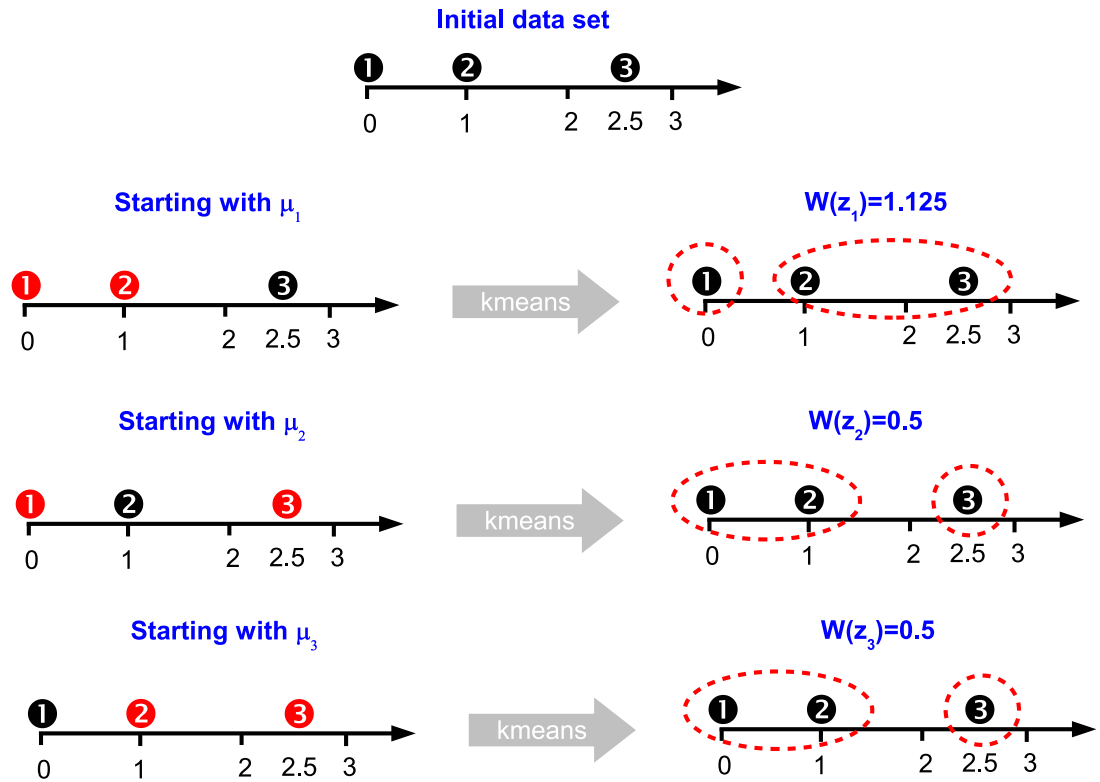
(a) Optimal clustering.



(b) Suboptimal clustering.

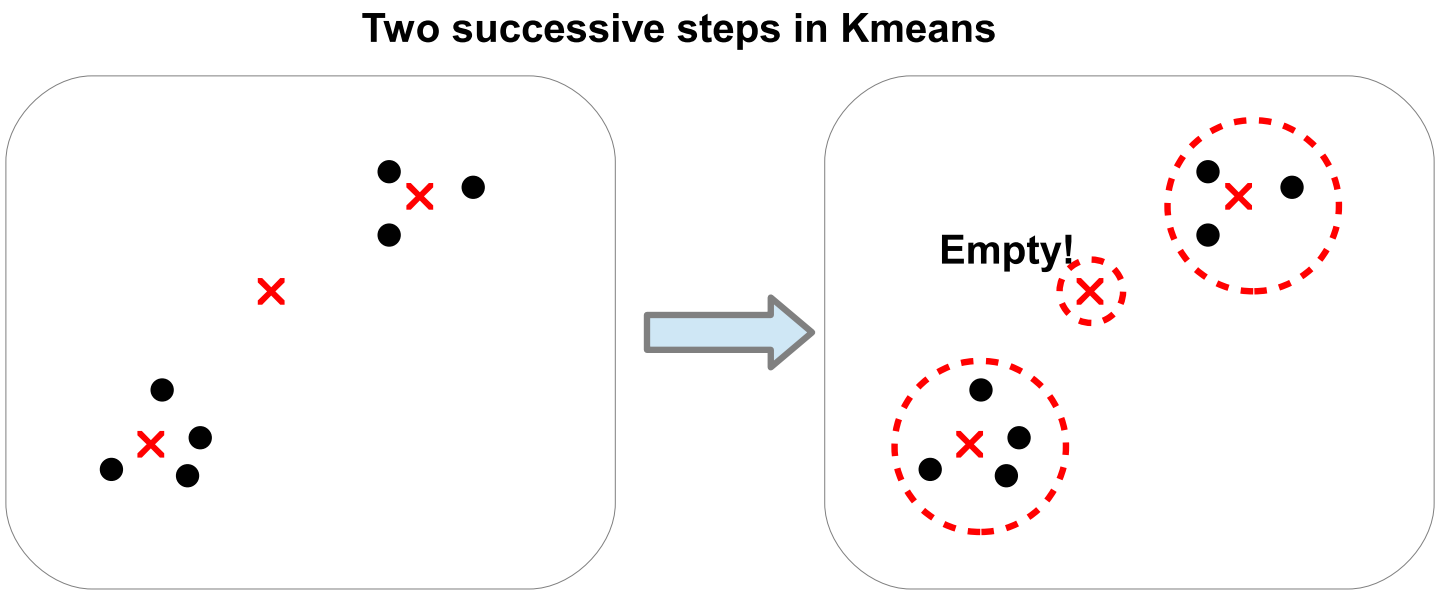
It is not a metric effect but a algorithm starting point effect

Local maxima with K-means: explanation



Run K-means from several random centers and keep the best W value

Empty clusters with K-means



Restart K-means when the empty cluster case occurs

What about hierarchical clustering?

- No problem of starting point
- No problem of local maxima
- But the price is strong constraints on nested partitions (see previous lesson)

Outline

1 Data factor

2 Dissimilarity factor (and co)

3 Algorithm factor

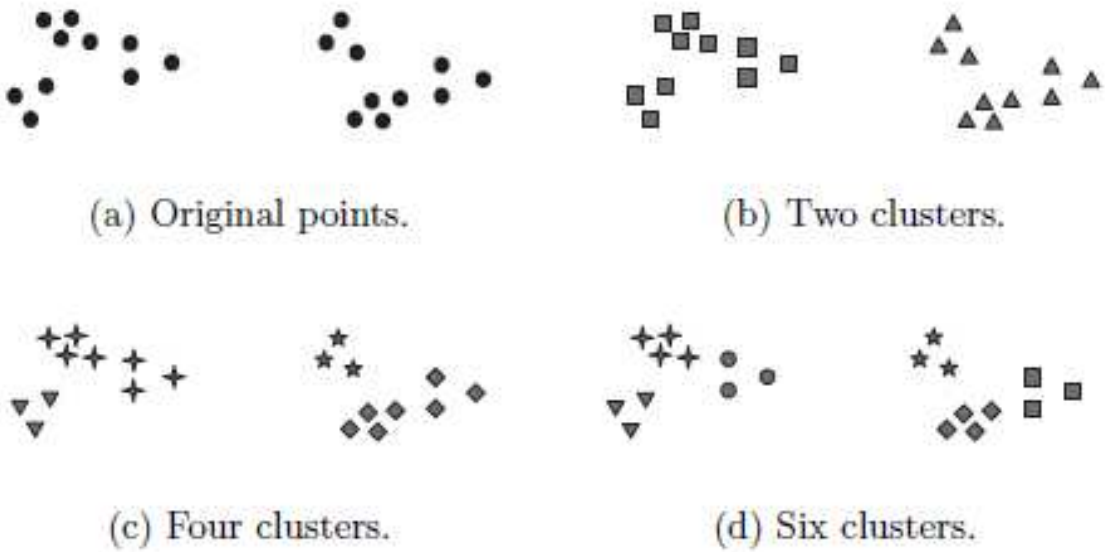
4 Number of clusters factor

5 User factor

6 To go further

Different values of K are valid!

[P.-N. Tan *at al.* (2005). Introduction to data mining, second edition, Addison-Wesley, Chap.8]



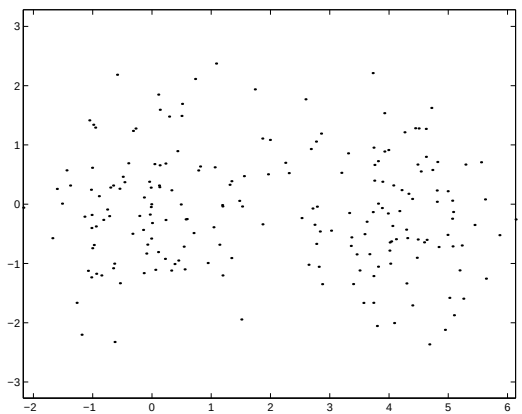
(a) Original points. (b) Two clusters.

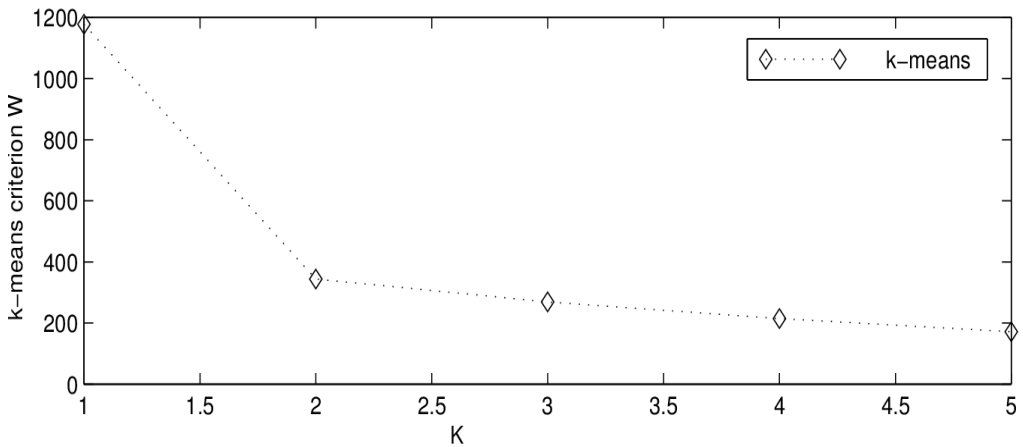
(c) Four clusters. (d) Six clusters.

Wk not enough

$$W_K = \arg \min_{z \text{ with } K \text{ clusters}} W(z)$$

- $W_{K+1} \leq W_K$
- $W_n = 0$





Some criteria to estimate K (1/3)

- The first "elbow" on the W_K curve [Hartigan, 1975]

$$\hat{K} = \min_K \left\{ K : \left[\frac{W(K)}{W(K+1)} - 1 \right] \times (n - K - 1) \geq 10 \right\}$$

- The Gap statistics measures the gap with uniformity [Tibshirani *et al.*, 2001]

$$\text{Gap}_K = \frac{1}{R} \sum_{r=1}^{R} \ln W_K^{(r)} - \ln W_K$$

with $W_K^{(r)}$ the within cluster sum of squares from a bth uniform data set with same range as the original data

$$\hat{K} = \min_K \left\{ K : \text{Gap}_K \geq \text{Gap}_{K+1} - \text{standard deviation} \left\{ \ln W_K^{(r)} \right\}_{r=1}^{R} \right\}$$

Some criteria to estimate K (2/3)

- Form of an ANOVA F-statistic [Calinski and Harabasz, 1974]

$$\hat{K} = \arg\max_K \frac{W_K/(K-1)}{B_K/(n-K)}$$

- Measure of how well the all $\boldsymbol{x}_i$ are clustered [Kaufman and Rousseeuv, 1990] :

$$\text{silouhette}(\boldsymbol{x}_i) = \frac{b_i - a_i}{\max(a_i, b_i)} \in [-1, 1]$$

  - a_i: average distance between $\boldsymbol{x}_i$ and all other observations of its clusters
  - b_i: average distance between $\boldsymbol{x}_i$ and points in the nearest cluster (minimizing b_i)

$$\hat{K} = \arg\max_K \frac{1}{n} \sum_{i=1}^{n} \text{silouhette}(\boldsymbol{x}_i)$$

Some criteria to estimate K (3/3)

- Possible high behaviour difference between criteria
- Expected since not the same point of view
- Where are theoretical guaranties? See next lesson...

[C. A. Sugar and G. M. James (2003). Finding the number of clusters in a data set: An information theoretic approach. Journal of the American Statistical Association.]

Simulation	Method	Cluster estimates									
		1	2	3	4	5	6	7	8	9	10
One (Five clusters, two dimensions)	ANOVA F-stat	0	0	0	0	98	0	1	1	0	0
	other	0	0	26	0	34	9	10	16	5	0
	Hartigan	0	0	0	0	0	0	1	5	18	76
	Silouette	0	51	21	4	24	0	0	0	0	0
	Gap	0	0	77	0	23	0	0	0	0	0
	other	0	0	3	4	92	0	0	0	1	0

Typology of methods for choosing the number of clusters[5]

- There exists many other empirical criteria (ex: cross-validation)
- There exists clustering methods including automatic choice of K (ex: DBSCAN)

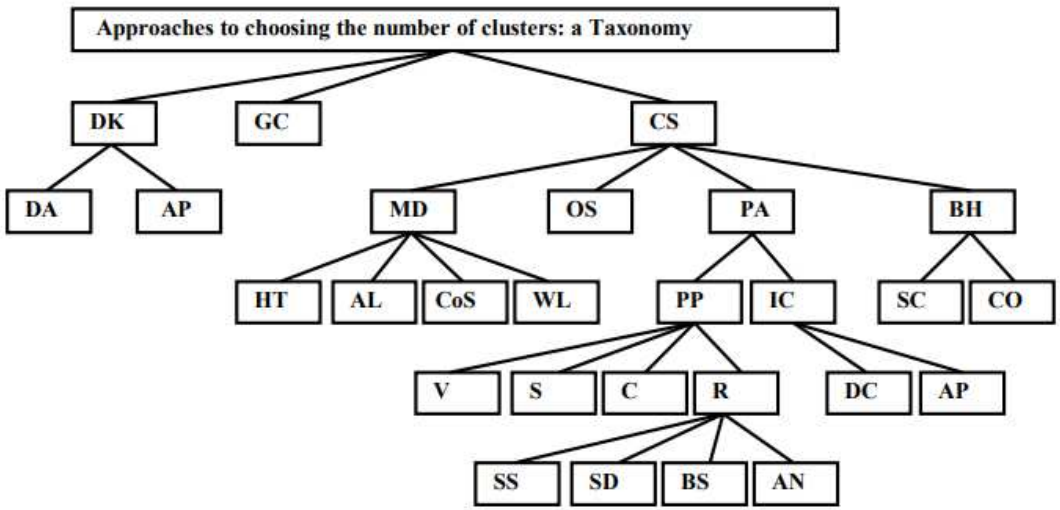


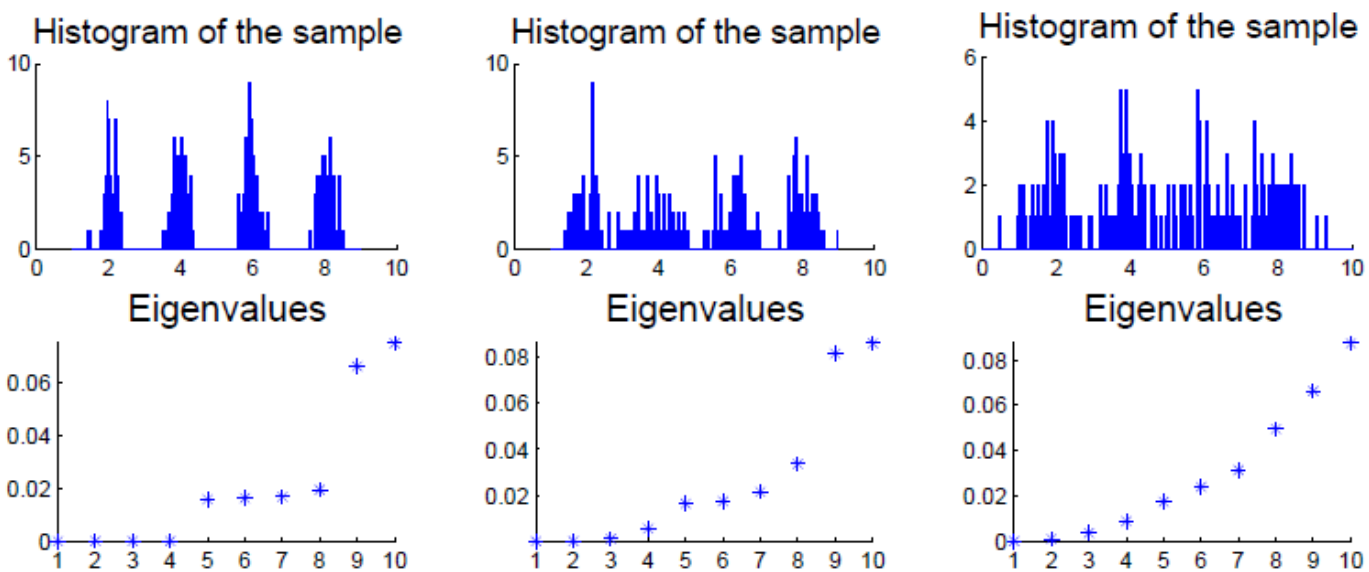
Figure 1: A taxonomy for the approaches to choosing the number of clusters as described in the paper. Here DK is Domain Knowledge in which DA is Direct Adjustment of algorithm and AP is Adjustment by Post-processing of clustering results, GC is modelling of the process of Generation of Clusters, and CS is Cluster Structures. This last item is further divided in MD, that is, Mixture of Distributions, P, Partitions, BH, Binary Hierarchies, and OS, Other cluster Structures. MD involves HT which is Hypothesis Testing, AL, Additional terms in the Likelihood criterion, CoS, Collateral Statistics, and WL, Weighted Likelihood; BH involves SC, Stopping according to a Criterion, and CO, using a cut-off level over a completed tree. Item PA covers PP, Partition Post-processing, and IC, pre-processing by Initialization of Centroids. The latter involves DC, Distant Centroids, and AP, Anomalous Patterns. PP involves V, Variance based approach, S, Structure based approach, C, Combining clusterings approach, and R, Resampling based approach. The latter is composed of SS, Sub-sampling, SD, Splitting the Data, BS, Bootstrapping, and AN, Adding Noise.

[5] Mirkin, Boris. (2011). Choosing the number of clusters. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery. 1. 252-260.

What about methods other than K-means?

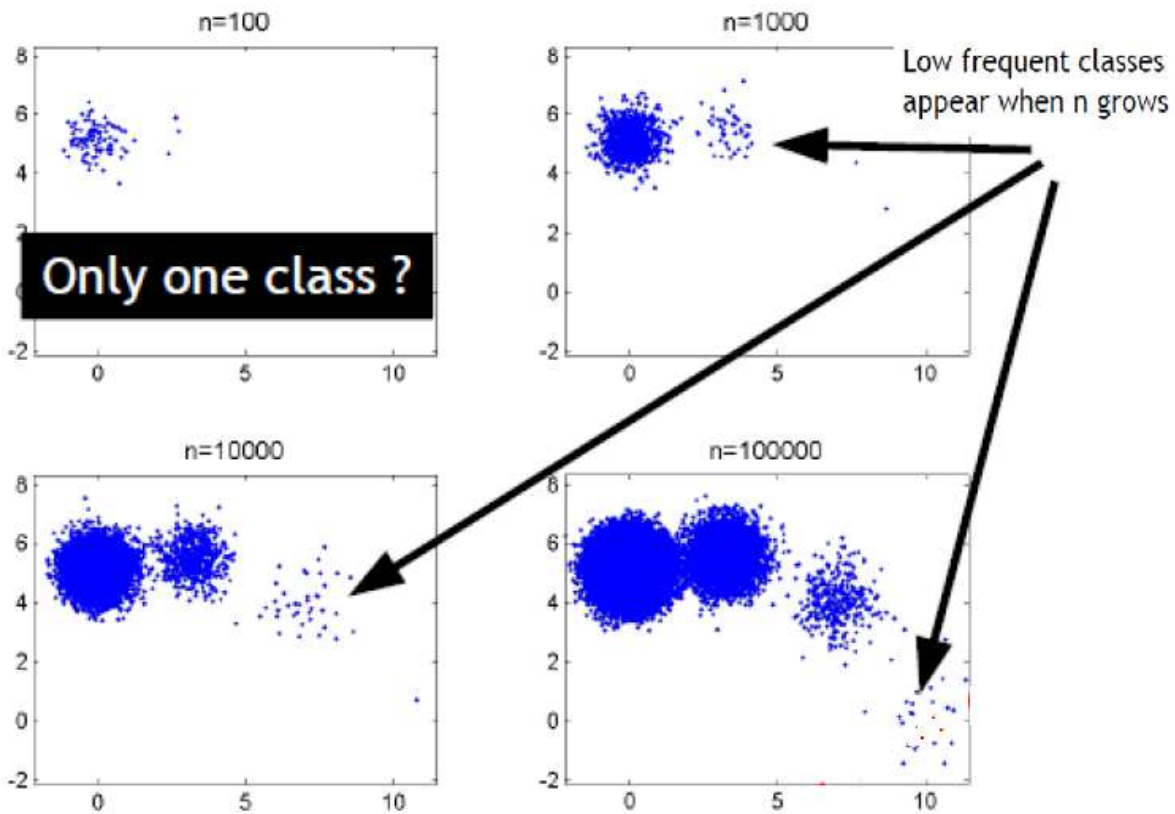
- Hierarchical clustering: previous criteria, an elbow in Δ...
- Spectral clustering: an elbow in the eigenvalues curve

[U. von Luxburg (2006). A Tutorial on Spectral Clustering.]



Number of clusters for large data sets

- When n increases, K is expected to do so
- See next lesson for formalizing that point

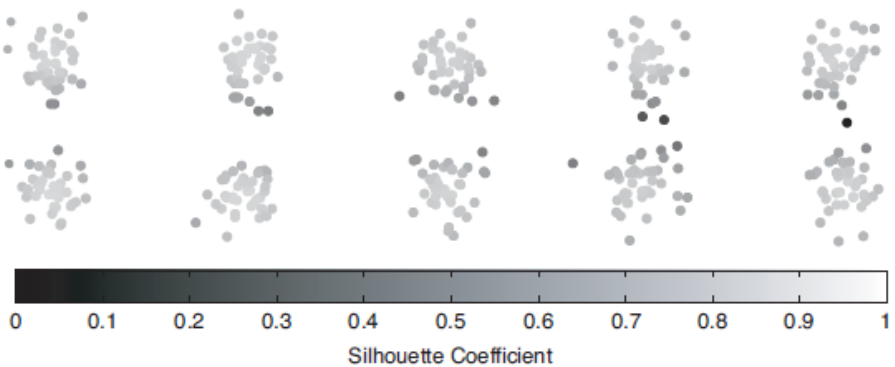


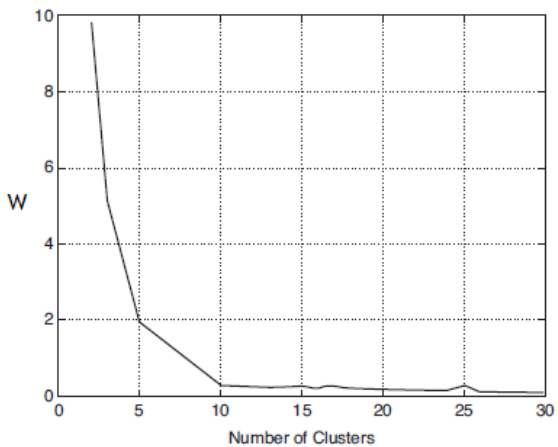
Gold rule

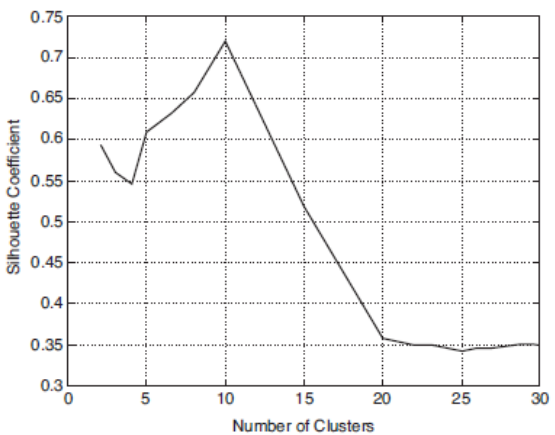
Retain a useful nb of clusters

- Some previous criteria are just here for guiding among a set of candidate K values
- Elbows are interesting for this task

[A. Jain *at al.*. Data Clustering: A Review.]







Outline

1 Data factor

2 Dissimilarity factor (and co)

3 Algorithm factor

4 Number of clusters factor

5 User factor

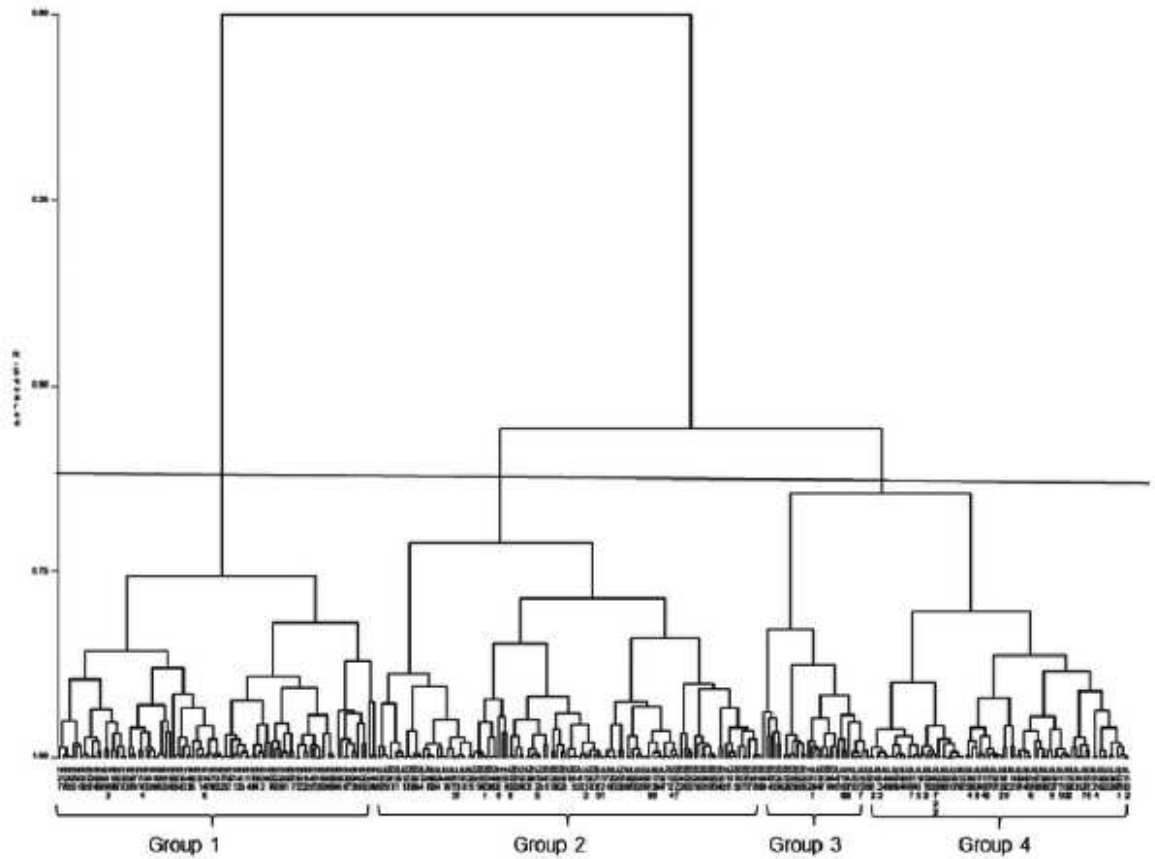
6 To go further

Partition and large data set

```
> z
   [1]  6  8  4  2  8  7  1  6  5  5 10  4  6  6  4  2  9  3  8  1  2  5  8  1  2 10  8  2  1  1  2  5  5  9  9  6  7  1  2  5  1  7 10  6  6  8  9  5  8  3  3  4  2
  [54] 10  6 10  2  9  4  8  9  3  8  7  8 10  9  8  7  5  6  7  8  6  8  4  3  8  2  3  8  2  3  7  7 10  8  7  6  1  7  5 10  4  9  5  8  1  7  4  6  5  6  9  2  6
 [107]  2  4  6  2  1  2  3  7  4  5  9  3  5  7  7 10 10  9  6  5  1  5  3  6  3  6  3  3  9  4  7  7  5  1  3  8  1  1  1  4  2  7  8  3  6  7  2  3  6  7 10  5  2
 [160]  6  9  9  4  8  6  4  9  2  6  7  3 10  5  1  3  1  7  4  8  2  5  9  6  4  3  6 10  9 10  9  5  7  7 10  8  2  1  5  6  2  1 10  7  5  6  5  2  8  5  5  3  6
 [213]  3  1  9  2  6  3  1  9  6  9  5  6  1  1  9  5  1  9  8  6  7  7  2  1  4  6  7  4  6  3  1  3  4  3  3  8  9  6  1  4  8  1  1 10  5  1  3 10  7  8  5  3  6
 [266]  3  2  9  8  7  9  7  6  7  9  8 10  8  2  5  7  4  9  2  9  2  7  7  1  4  9  6  4  3  4  3  9  1 10  8  4  1  6  2  6  4  4  7  6  2  5  2  5  8  6  2  1  7
 [319]  6  6  8  3  2  4  9  7  2  4  1 10  5 10  3  7  9  3  2  8  8  2  7  8  2 10  1 10  5  7  2  8  9  5  8  4  4  4  7  7  6  4 10  7  9  3  9  3  8  1  6  6 10
 [372]  7  8 10  4  9  2  6  4  7  8  6  5  4  2  1  5  8  9  1  3  5  7  2  1 10 10  9  4  1  9  9  6  7  1  2  9 10 10  9  4  8  9  2  2 10  9  1 10  6  6 10  2  4
 [425]  1  4  3  5  7  3  2  8 10  3  9  9  3  3  7  5 10  8  1  9 10  1  5  3  7  2 10  8  7 10  9  3  6  6  7  2  7  6  6  2  3  8  5  5  5  1  7  1  1  7  3  5  9
 [478]  5  3  7  2  3  7  7  5  3  5  3 10  6  9 10  9  3  7  8  3  8  5  4  6 10  6  2 10  2  4 10  2 10  3  2  2  2  7  5  2  5  4  5  7  1  4 10  9  3  1 10  7  3
 [531]  7 10 10  9  2  9  1  9  8  6  2  6 10  4  1  7 10  3  5  7  6  8 10  5  1  2  4  6  3  8  4 10  6  9  1  1  5  8  4  6  2  9  9  6  1  8  8  4  6  7  8  3  8
 [584]  3  5  4  8 10  6  7  1 10 10  3  4  5  4  5  7  9  7  1  2  9  1  1  2  1  7  8  3 10  8  9  3  2  9  8  9  7  9  9  6  3 10  3  8  9  2 10  9  9  5  3  6  1
 [637]  2  1  6 10  4  4  6  4  3  2  2  4 10  9  6  6  4 10  4 10  1 10  5  3  2  9  4  1  2  6  4  3  2  5  2  3  6  2  4  7 10  1  3  8  5  2  1  3  1  9  1  1  9
 [690]  9  2  3  7  3  6  8  6  3  8  4  5  9 10  2  7  5  2  3  6  7  7  6  6  4  7  6  9 10  7  4  9  7  3  1  8  9  3  4 10 10  6  2  3  2  7  4  4  8  1  8  9 10
 [743]  6  7  3 10  4  3  2  7  2  2  8  6  9 10  8  8  6  3  1  1  2 10  4  1  9  1 10  2 10  8  8 10  6  9  5  6  6  2  1  7  6  9  9  4  3  1  1  8  4  7  5  8  1
 [796]  5  4  9  2  8  2  9  3  6  5  3  1  3  8  5  8  5  7  2  7  1  6  1  4  4 10  2  8  3  1  4  5  9  9  3 10  2  3 10  9  2  3  4  6  5 10  6  8  2  7  4  3  9
 [849]  6  2 10  2  7  5  2  9  7  6  3  3  1  7  9  4  9 10  6  1  4  9  1  1  2  6  2  2  7  9  8  3 10  7  4  3  7  9  4  2  9  1 10 10  2  7  6  8  2 10  4 10 10
 [902]  6  7  1  9  3  8  8 10  4  9  5  7  9  1  9  5 10  8  7  8  5  1 10  3  9  1 10  1 10 10  4 10 10  5  5  3  3 10  1  7  3  4  1  7  4  9  4  1  9  9  8  2  8
 [955]  1  4  5  1  2  6  2  7  5  2 10  2  5  1  7  6  5  9  4  8 10  7  3  4  5  9  9  8  3  9  6  9  1  4  4  6  4  5  8  8  6  5  4  9  2  3  5  4  2  9  4  4  2
[1008]  7  8  5  1  9  3  1  4  8  1  7  7  1  4  9  4  1  5 10  1  3  6  2  4  1  5  9  6  5  7  4  1  1  8  8  7  3 10  2  9  5 10  8  2  4  9  3  9  2  7  5  5  2
[1061]  2  7  5  1  8  2  6  9  1  7  3  2  1  1  8  8  1  6  7  1  4 10  2  1  4  2  8  8  3  9  6  3  7  1 10  5  8  5  1  3  7  5  1  9  6 10  8  2  5  5  5  5  6
[1114]  9  2  4  1  2  9  1  9 10  9  5  9  9  6  6  1  9 10  3  1  7 10  3  9  9  4  8  5  2  1  9  5  8  5  2  7  1  7  5  2  3  2  6  8  6  1  7  2  9 10  1  9  7
[1167]  5  5  5 10  4  1  2  6  4  9  3  4  9  7  4  9  6 10  7  7  3  4 10  3  3  9  5  4  3  6  3  8  9  4  9  9 10  9  4  9  2  9  7  7  6  5  3  4  3 10  9 10  8
[1220]  7  3  2  9  8  8  5  8  3  6  2 10  6  9  5  7  9  5  5  6  1  7  7  3  3  7  1 10  9  9  5  9  7  1  9  9  6  5  8  6  5  4 10  2  2  6  8  5  4 10  6  9  5
[1273]  9  9  8 10  7  6 10  8  2  9  7  9  8  2  8 10  9  2  2  1  7  5  3  9  7  8  6  4  2  3  6  3  8  4 10  8  4  8  4 10  1  1  7  9  3  1  1  7  7  2  6  1  4
[1326]  2  2  5  3  2  1  4  5  3 10  8 10  8  3  3  1  4  3  4  1  7  8  6  2  8  3 10  6 10  2 10  3 10  9  8  7  6  2  6  5 10  8  2 10  2  7  6  6  6  2  9  4  4
[1379]  2  2  2  9  7  1  7  3 10  3  1  2 10  3  1  7  4  3  1  5  1  1  6  4  6  8  2  4 10  8  2  5  4  2  9  2  5  3  4  4  3  2  8  7  3  8  3  5  1  5  9  7  6
[1432]  5  1  8  6  2  1  2  3  5 10  7  6  4  4  6  1  4 10  7  9  8  8  7  4  5  2  5  9 10  6  8 10  1  3  5  1  2  7  4  3  4  6  5  7  1  2  9 10  6  4  2  7 10
[1485]  5  3  2 10  1  4  9  5  4  9  4  5  3  5  9  9
```

Little readable. . .

Dendrogram and large data set



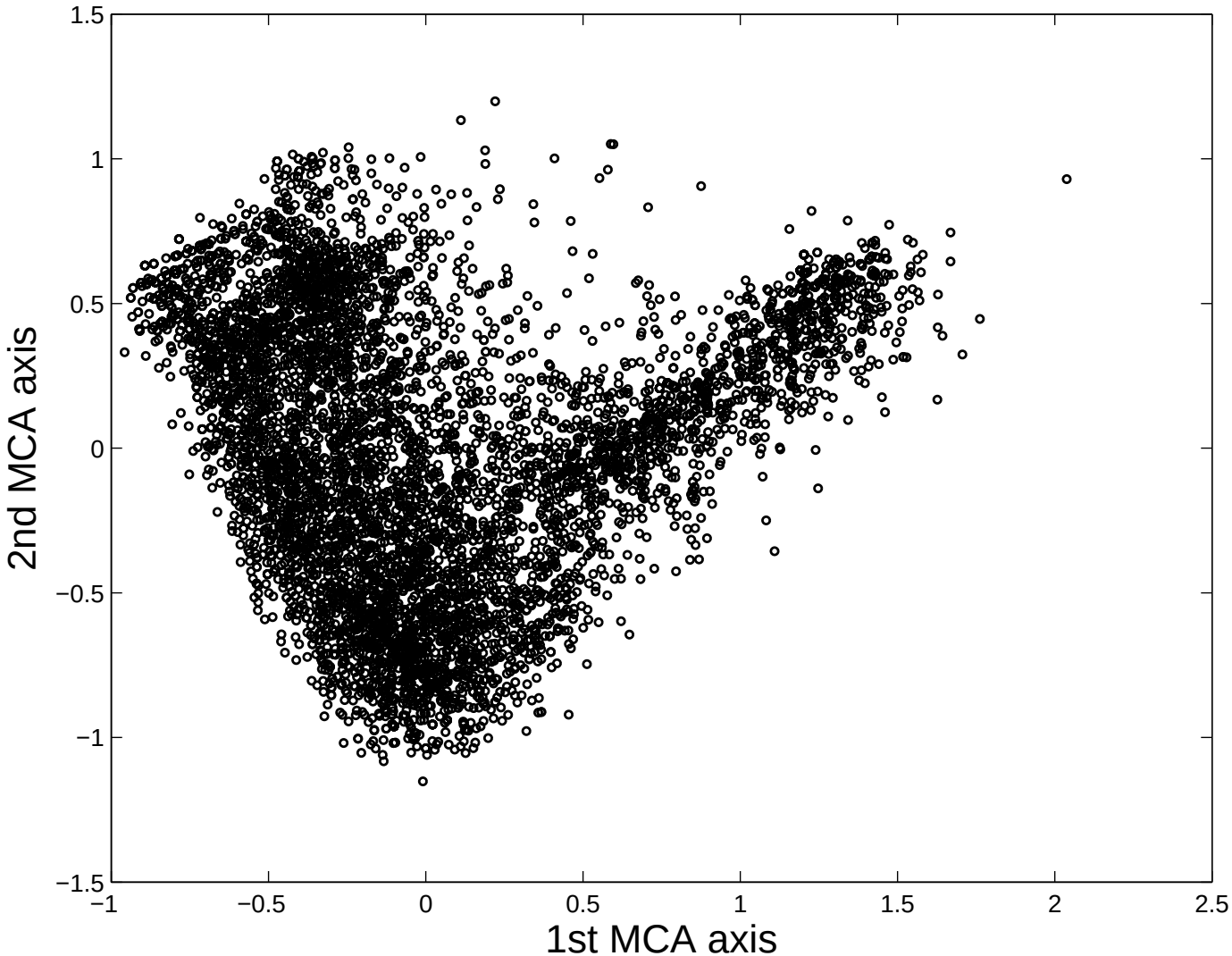
Little readable. . .

Marketing Data: description

- $n = 6876$ households of the San Francisco bay
- $d = 13$ categorical variables

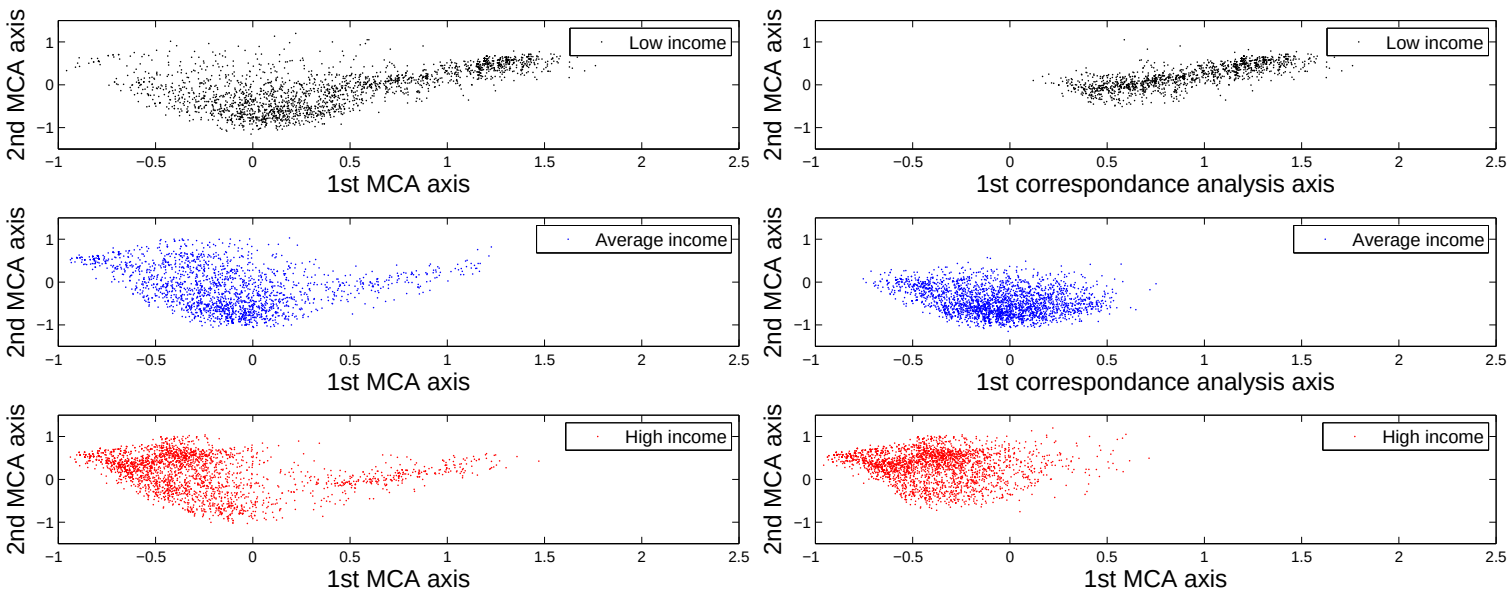
1. SEX: 1. Male 2. Female
2. MARITAL STATUS: 1. Married 2. Living together, not married 3. Divorced or separated 4. Widowed 5. Single, never married
3. AGE : 1. 14 thru 17 2. 18 thru 24 3. 25 thru 34 4. 35 thru 44
4. EDUCATION: 1. Grade 8 or less 2. Grades 9 to 11 3. Graduated high school 4. 1 to 3 years of college 5. College graduate 6. Grad Study
5. OCCUPATION: 1. Professional/Managerial 2. Sales Worker 3. Factory Worker/Laborer/Driver 4. Clerical/Service Worker 5. Homemaker 6. Student, HS or College 7. Military 8. Retired 9. Unemployed
6. HOW LONG HAVE YOU LIVED IN THE SAN FRAN./OAKLAND/SAN JOSE AREA? 1. Less than one year 2. One to three years 3. Four to six years 4. Seven to ten years 5. More than ten years
7. DUAL INCOMES (IF MARRIED): 1. Not Married 2. Yes 3. No
8. PERSONS IN YOUR HOUSEHOLD: 1. One 2. Two 3. Three 4. Four 5. Five 6. Six 7. Seven 8. Eight 9. Nine or more
9. PERSONS IN HOUSEHOLD UNDER 18: 0. None 1. One 2. Two 3. Three 4. Four 5. Five 6. Six 7. Seven 8. Eight 9. Nine or more
10. HOUSEHOLDER STATUS: 1. Own 2. Rent 3. Live with Parents/Family
11. TYPE OF HOME: 1. House 2. Condominium 3. Apartment 4. Mobile Home 5. Other
12. ETHNIC CLASSIFICATION: 1. American Indian 2. Asian 3. Black 4. East Indian 5. Hispanic 6. Pacific Islander 7. White 8. Other
13. WHAT LANGUAGE IS SPOKEN MOST OFTEN IN YOUR HOME? 1. English 2. Spanish 3. Other

Marketing Data: MCA visualization



Marketing Data: partition overview

err=46%	$\hat{z}$		
z	$-19999\$$	between	$+40000\$$
$-19999\$$	1001	166	282
between	996	1023	624
$+40000\$$	292	802	1690



true partition

estimated partition

Marketing Data: cluster description

Cluster proportion

Income	Low	Average	High
π_k	0.4036	0.3855	0.2109

Marital status

Income	Married	Living together, not married	Divorced or separated	Widowed	Single, never married
Low	0.0037	0.0253	0.0096	0.0000	0.9613
Average	0.0035	0.1364	0.2486	0.0762	0.5353
High	0.9504	0.0496	0.0000	0.0000	0.0000

Householder status

Income	Own	Rent	Live with Parents/Family
Low	0.0548	0.0811	0.8641
Average	0.2493	0.7011	0.0496
High	0.6644	0.3264	0.0091

etc.

SPAM E-mail database: description[7]

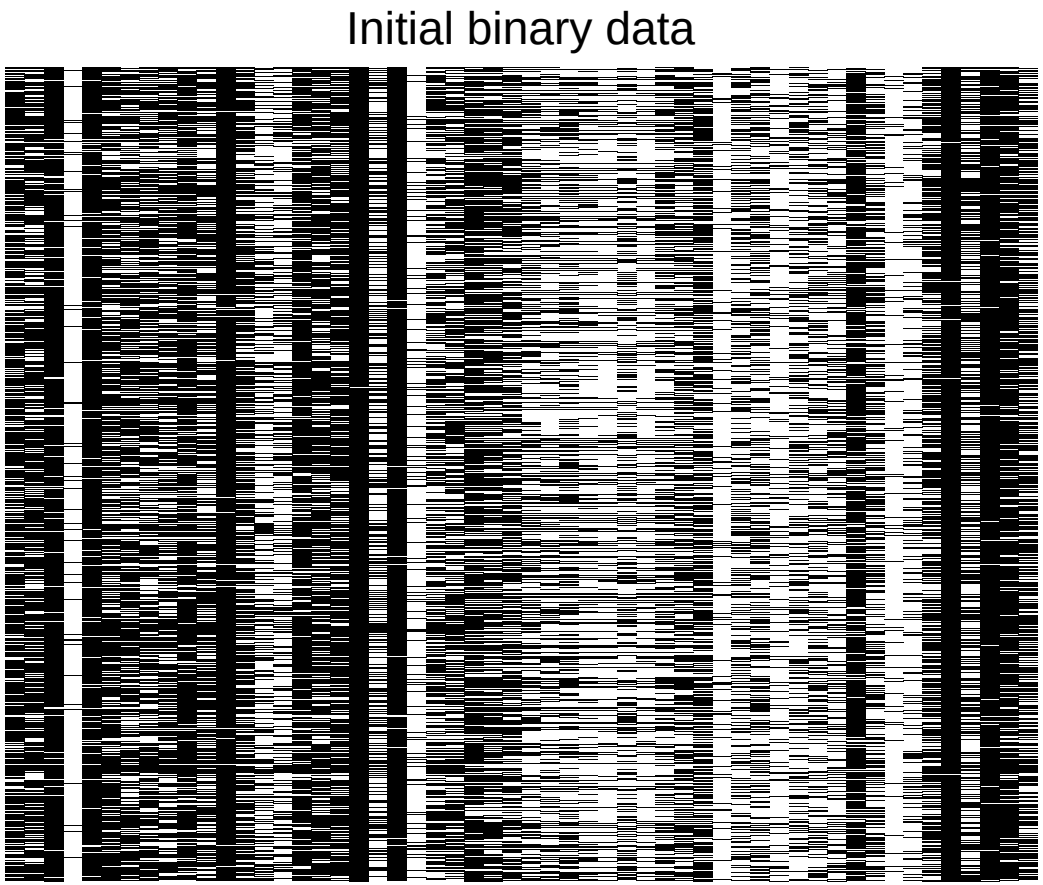
- $n = 4601$ e-mails composed by 1813 "spams" and 2788 "good e-mails"
- $d = 48 + 6 = 54$ continuous descriptors[6]
  - 48 percentages that a given word appears in an e-mail ("make", "you"...)
  - 6 percentages that a given char appears in an e-mail (";", "$"...)
- Transformation of continuous descriptors into binary descriptors

$$x_{ij} = \begin{cases} 1 & \text{if word/char } j \text{ appears in e-mail } i \\ 0 & \text{otherwise} \end{cases}$$

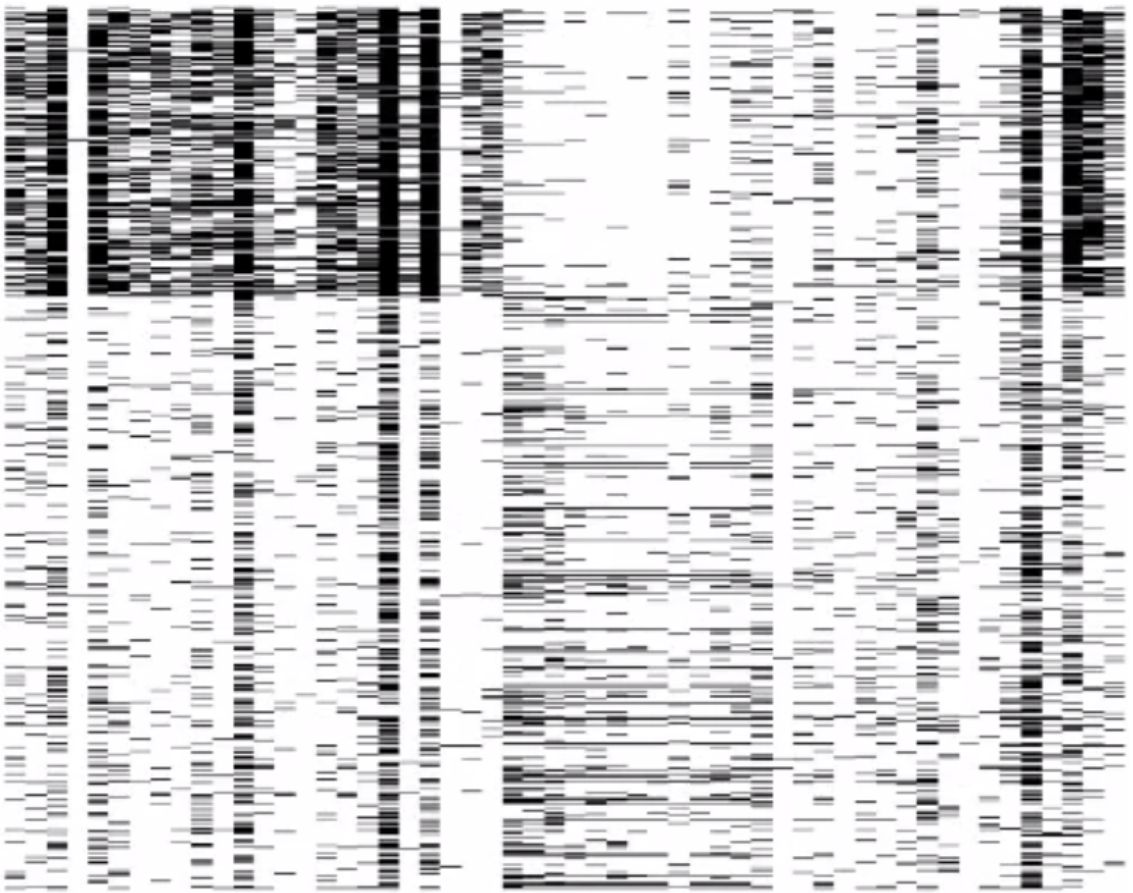
[6] There are 3 other continuous descriptors we do not use

[7] https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/

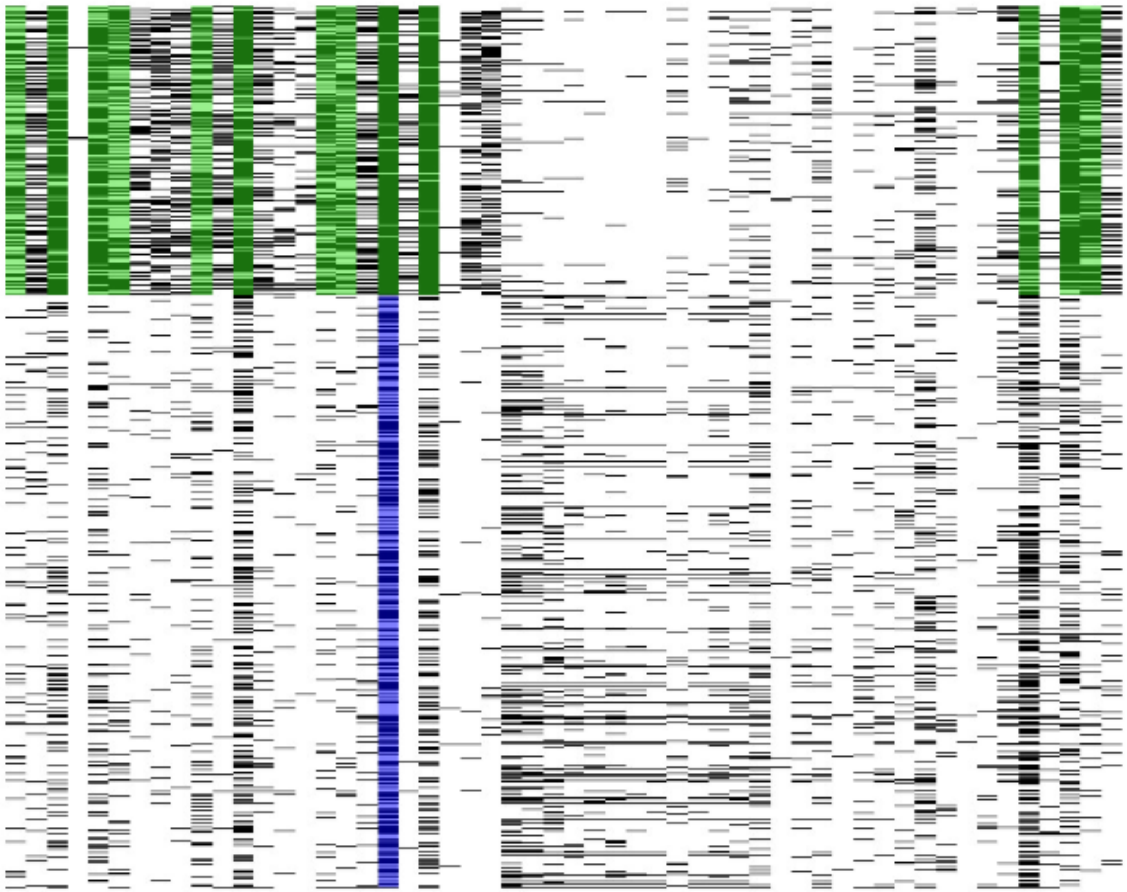
SPAM E-mail database: raw visualization



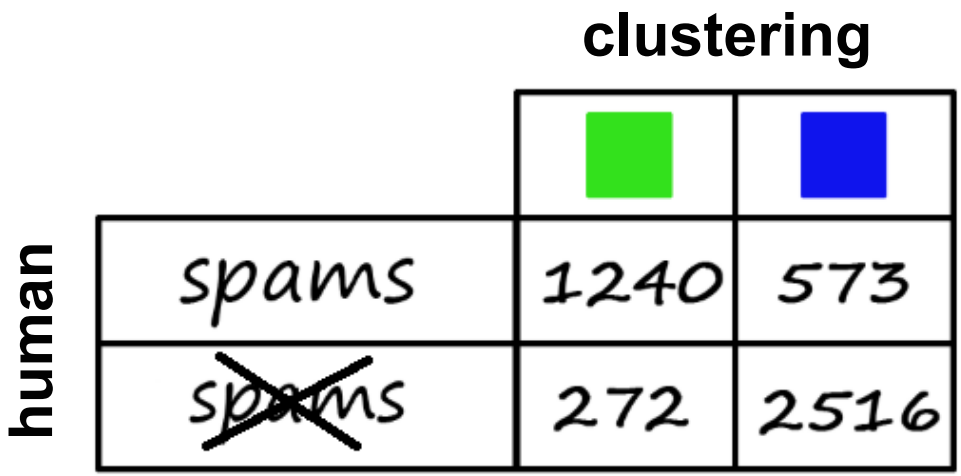
SPAM E-mail database: two clusters (1/4)



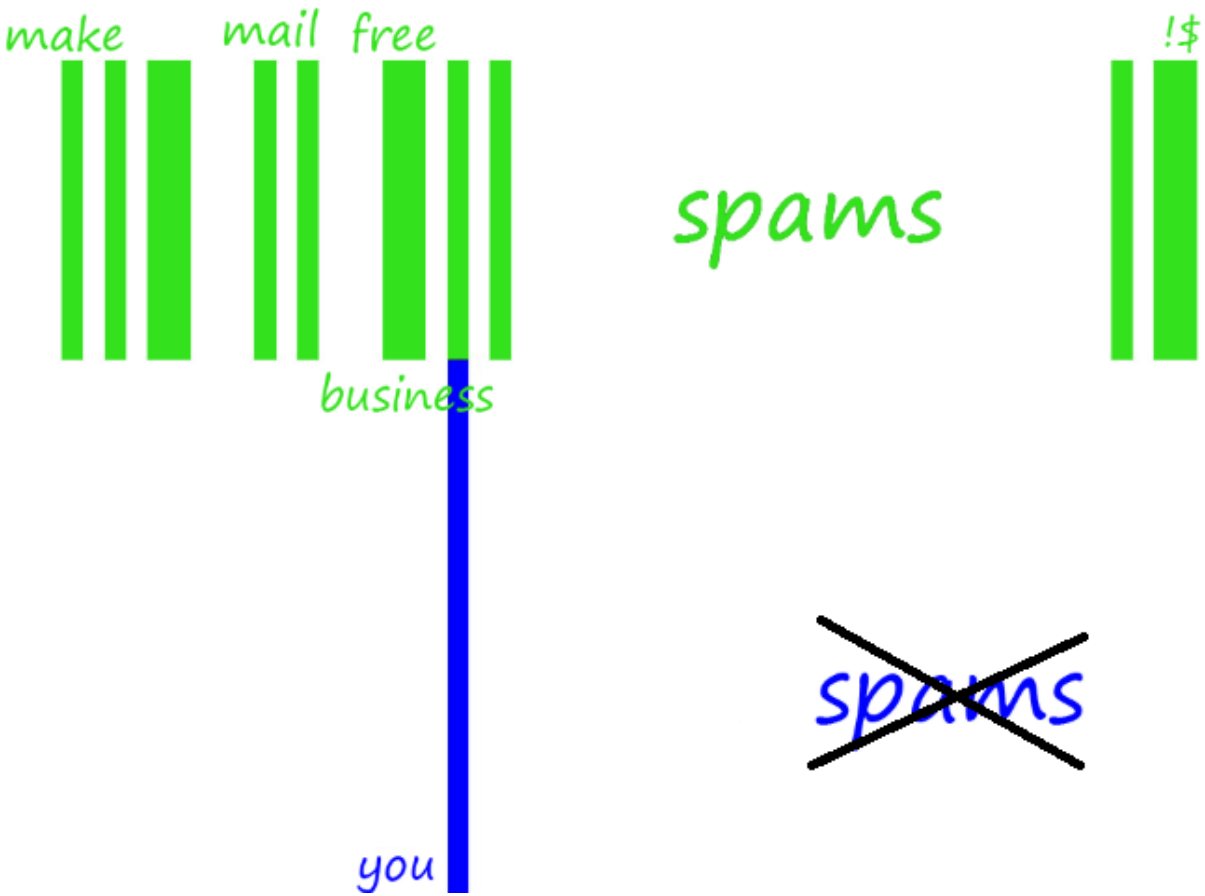
SPAM E-mail database: two clusters (2/4)



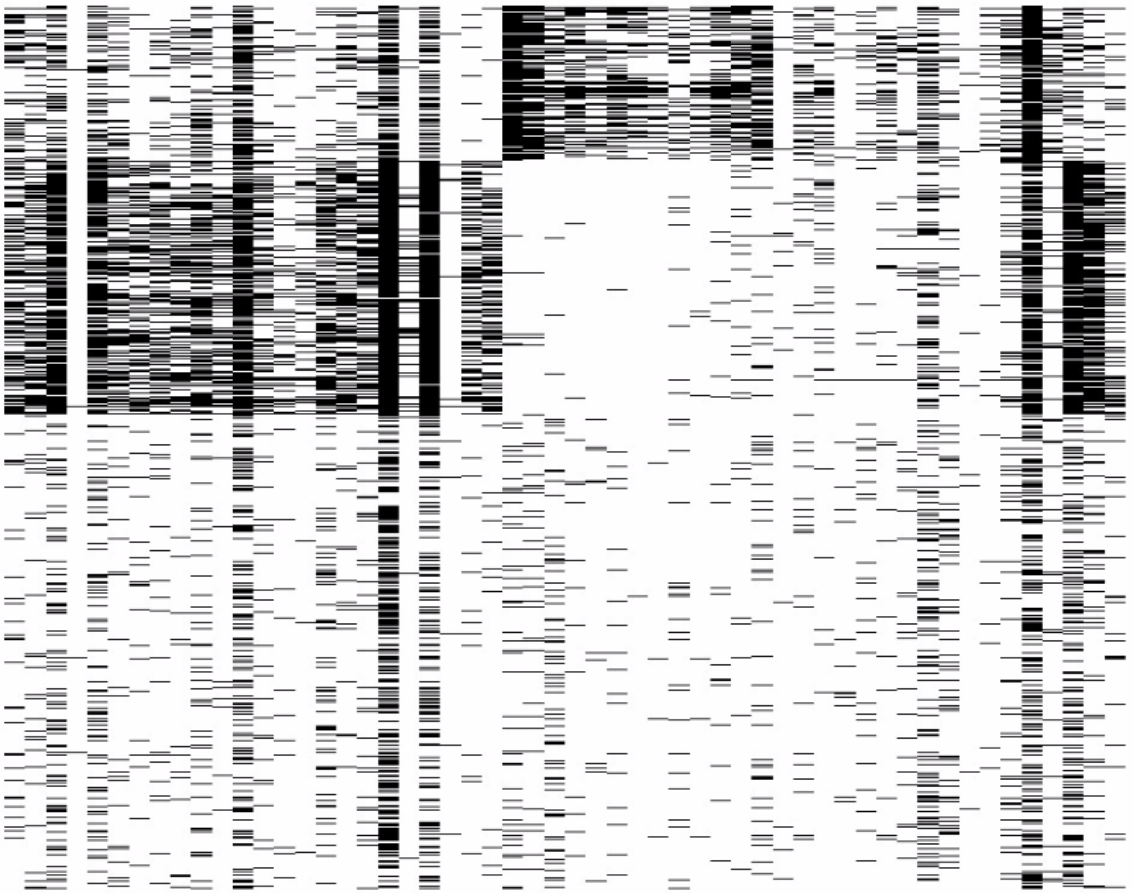
SPAM E-mail database: two clusters (3/4)



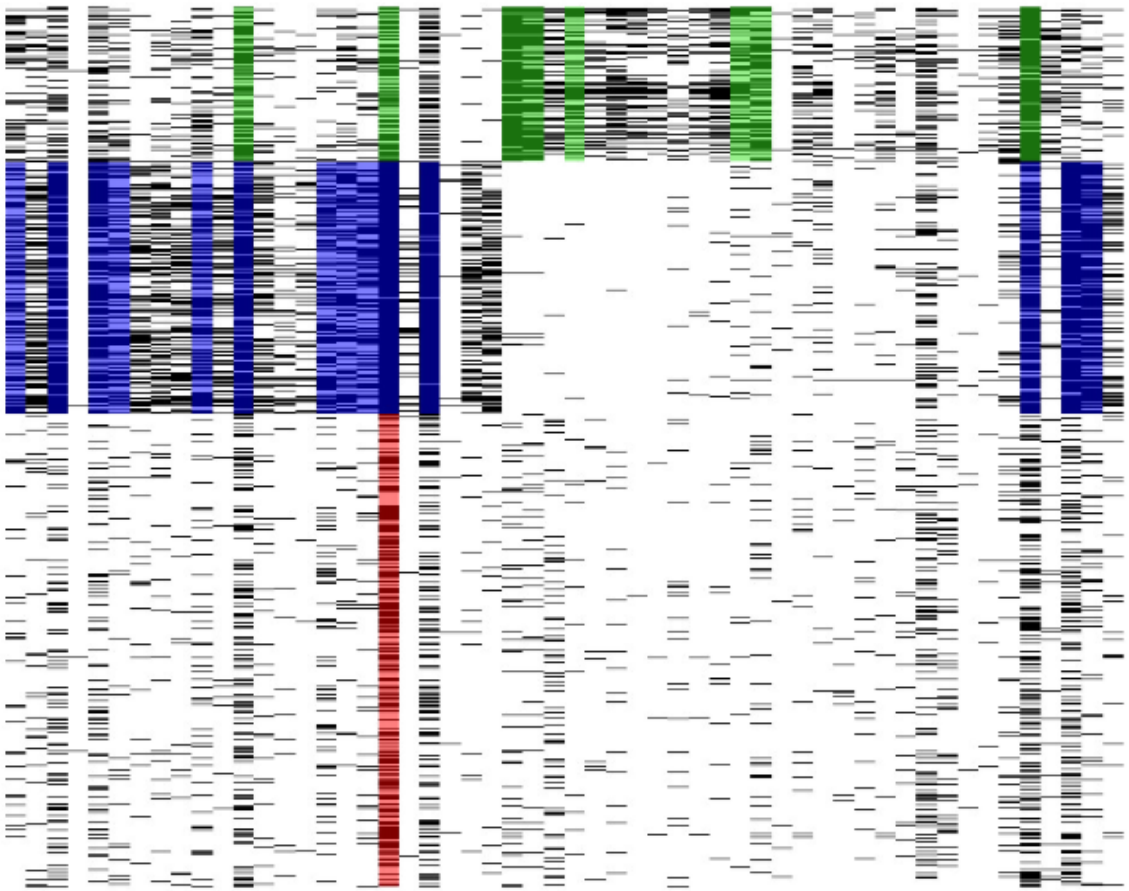
SPAM E-mail database: two clusters (4/4)



SPAM E-mail database: three clusters (1/4)

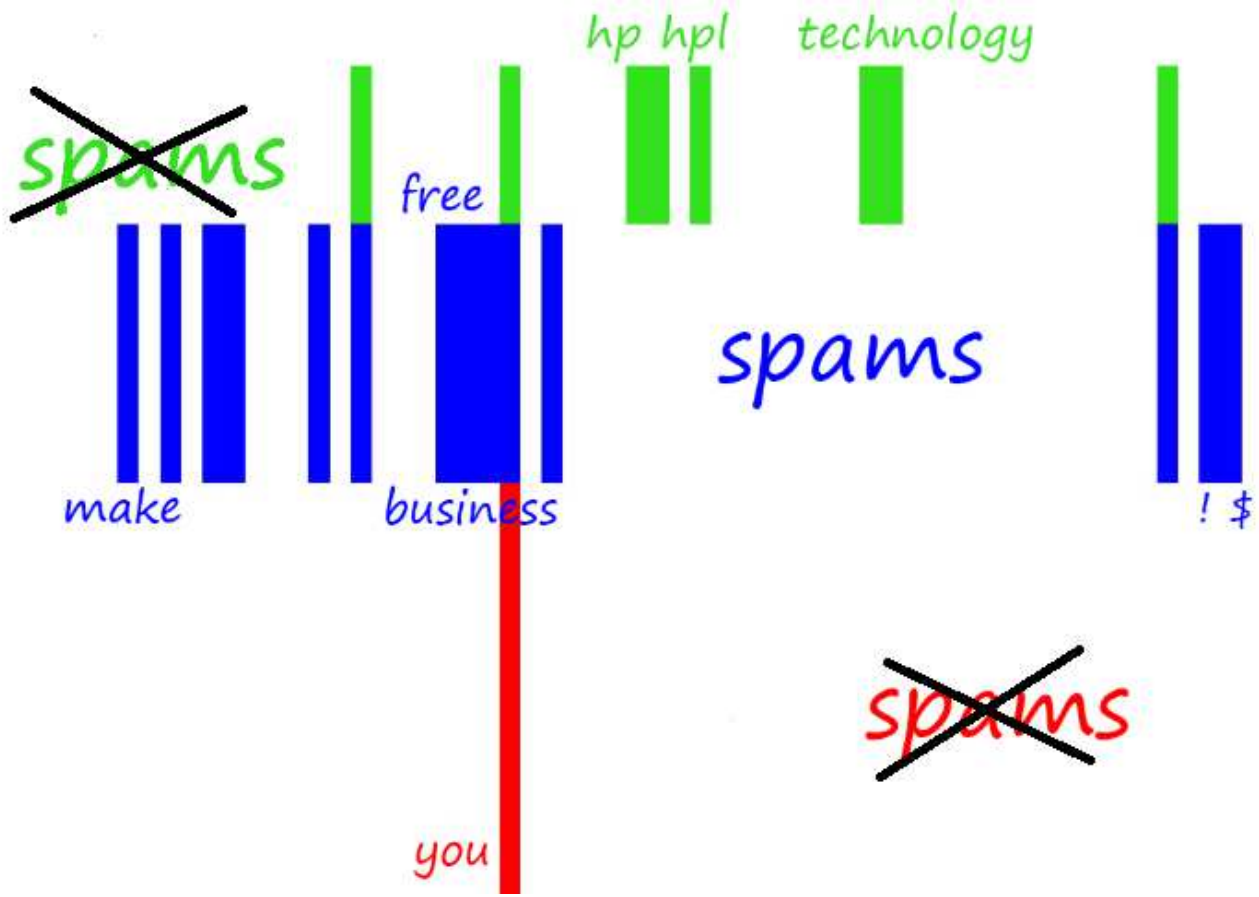


SPAM E-mail database: three clusters (2/4)



SPAM E-mail database: three clusters (3/4)

SPAM E-mail database: three clusters (4/4)



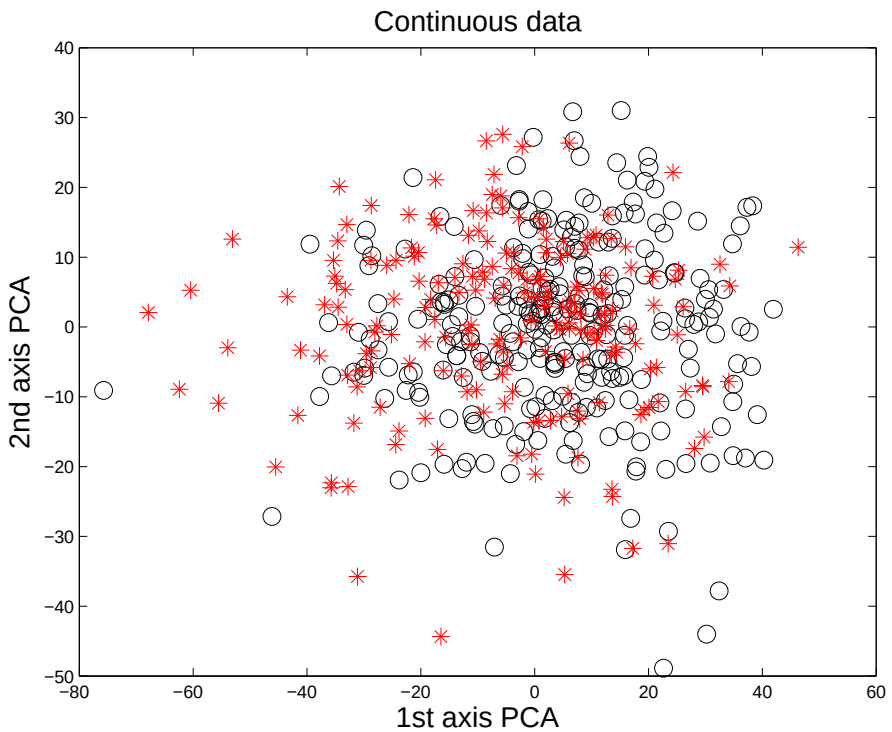
Prostate cancer data: description[8]

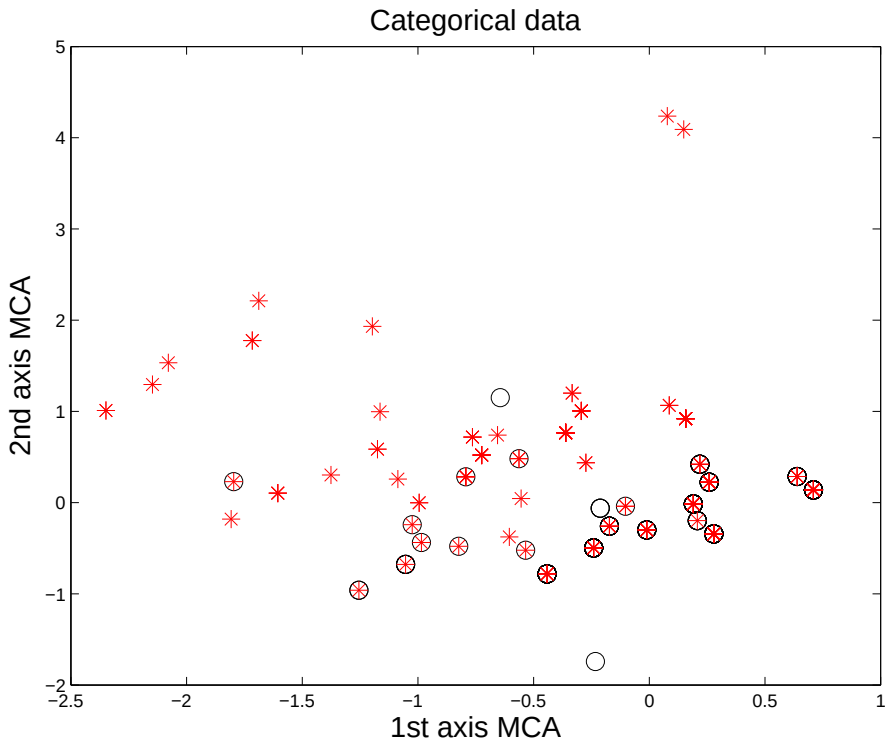
- Individuals: 506 patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease
- Variables: $d = 12$ pre-trial variates were measured on each patient, composed by eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histolic grade, serum prostatic acid phosphatase) and four categorical variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases)
- Some missing data: 62 missing values ($\approx 1\%$)

We forget the classes (Stages of the desease) for performing clustering

[8] Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488

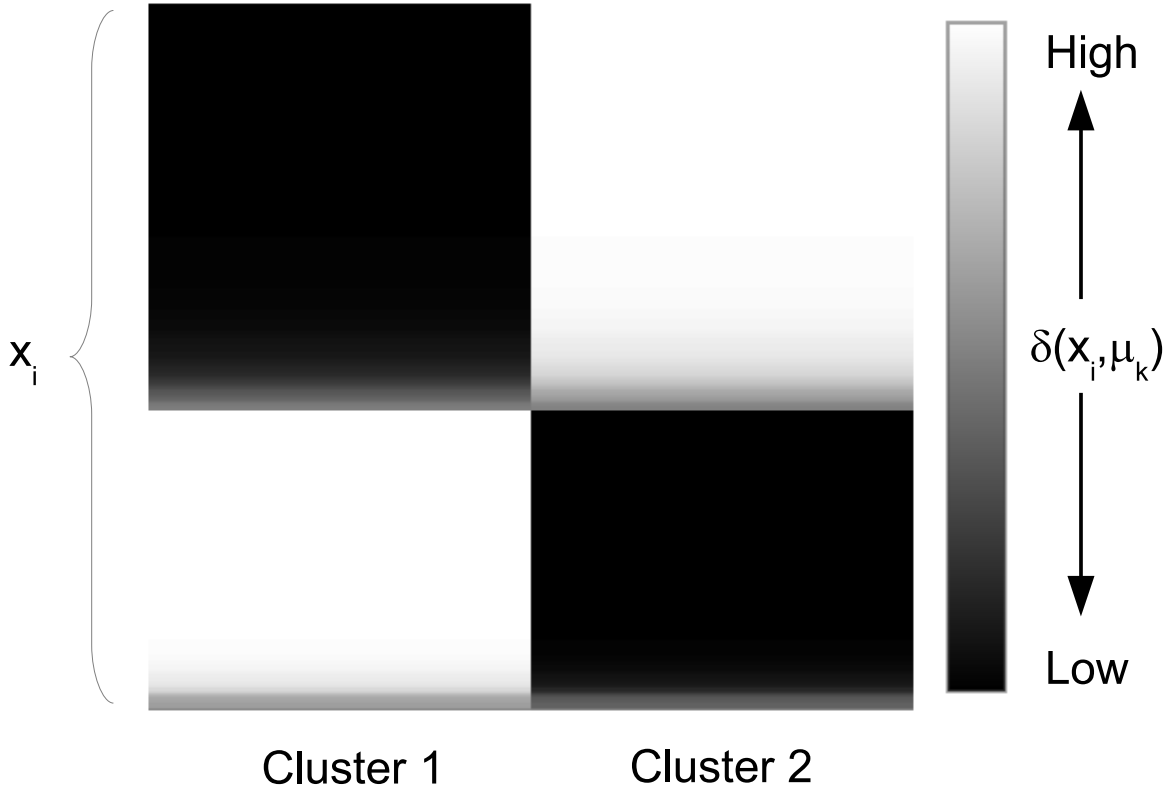
Prostate cancer data: PCA and MCA partition visualization





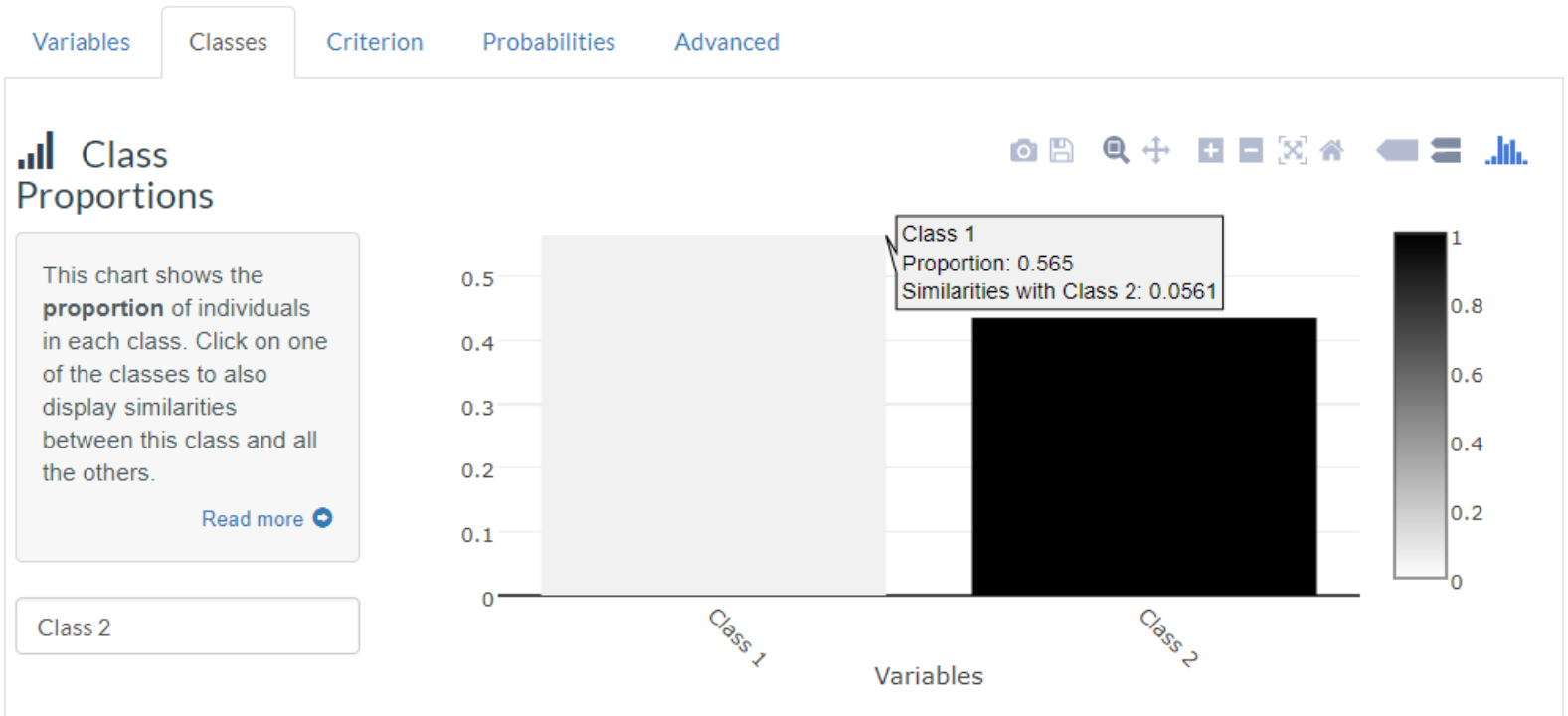
Seems to be not well separated

Prostate cancer data: chart of individuals sorted by distance to centers visualization



In fact it is well separated

Prostate cancer data: cluster weight

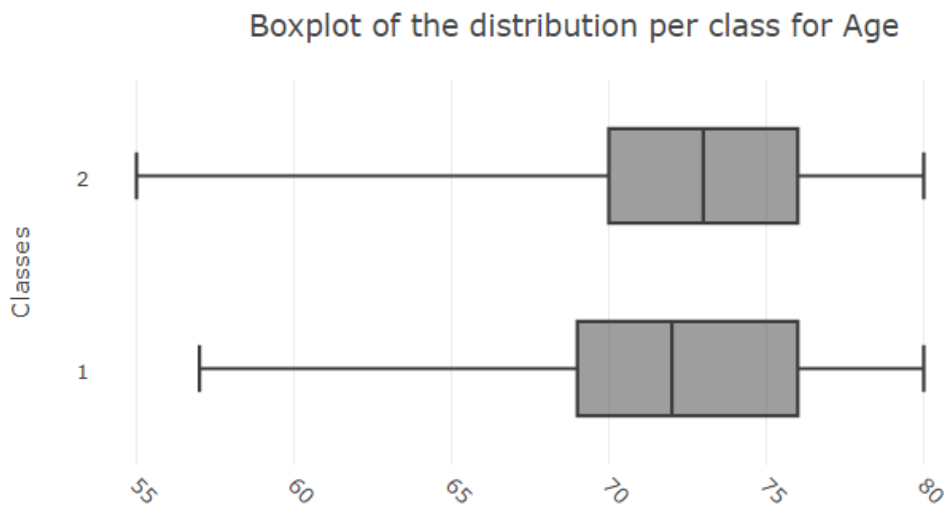


Prostate cancer data: variable "Age" difference between clusters

Variable Parameters

This chart summarizes the distribution of the selected variable.

Age



Age (Gaussian)

Hide model parameters

Class 1	Class 2
mean: 71.534, sigma: 6.760	mean: 71.313, sigma: 7.463

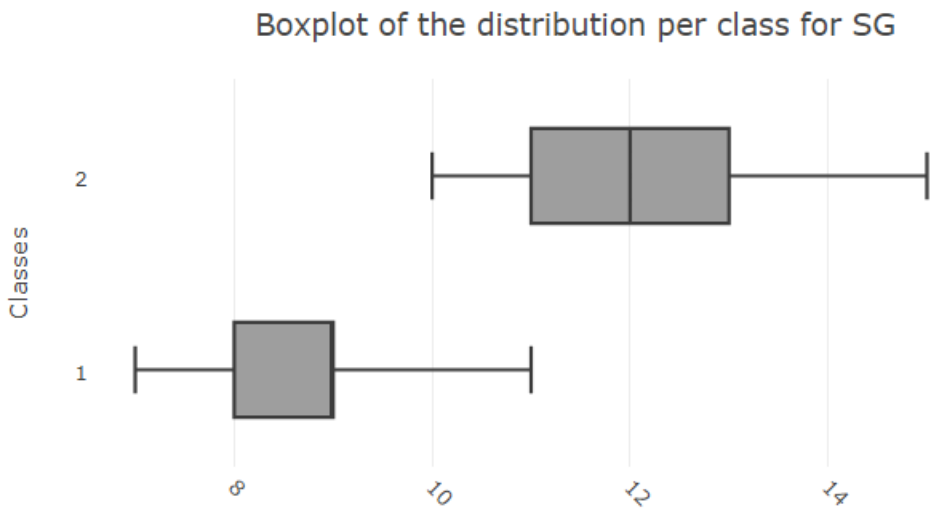
"Age" seems to be not very discriminant

Prostate cancer data: variable "SG" difference between clusters

Variable Parameters

This chart summarizes the distribution of the selected variable.

SG



SG (Gaussian)

Hide model parameters

Class 1	Class 2
mean: 8.940, sigma: 1.154	mean: 12.087, sigma: 1.405

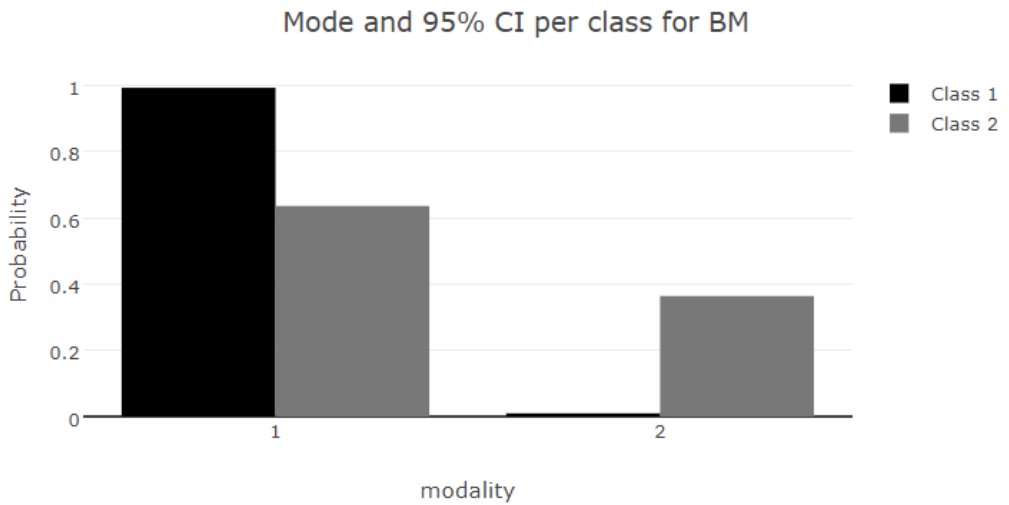
"Age" seems to be very discriminant

Prostate cancer data: variable "BM" difference between clusters

Variable Parameters

This chart summarizes the distribution of the selected variable.

BM



BM (Multinomial)

Hide model parameters

Class 1	Class 2
scatter: [0.993,0.007]	scatter: [0.633,0.367]

"BM" seems to be very discriminant

Outline

1 Data factor

2 Dissimilarity factor (and co)

3 Algorithm factor

4 Number of clusters factor

5 User factor

6 To go further

Next lesson

To go further towards clustering evaluation, there is a need to further formalize...

Introduction to cluster analysis and classification:
Formalizing clustering